PA Council
Report NR 22

David Magnusson

A STUDY OF RATINGS BASED ON TAT

A STUDY OF RATINGS BASED ON TAT

BY

DAVID MAGNUSSON
FIL. LIC.

AKADEMISK AVHANDLING

SOM MED VEDERBÖRLIGT TILLSTÅND AV HUMANISTISKA FAKULTETEN VID STOCKHOLMS HÖGSKOLA FÖR VINNANDE AV FILOSOFISK DOKTORSGRAD KOMMER ATT TILL OFFENTLIG GRANSKNING FRAMLÄGGAS MÅNDAGEN DEN 11 MAJ 1959 KL. 10 F. M. Å PSYKOLOGISKA INSTITUTET, TEKNOLOGGATAN 8

A STUDY OF RATINGS BASED ON TAT

BY

DAVID MAGNUSSON

THE SWEDISH COUNCIL FOR PERSONNEL ADMINISTRATION
HANTVERKARGATAN 78 — STOCKHOLM — SWEDEN

GRAFISKA KONSTANSTALTEN TRYCKERI AB

STOCKHOLM 1959

Foreword

One of the most important tasks of the personnel administration department of an undertaking is the recruiting of new staff. The selection is often influenced by other factors than the merits of the applicants; general standards of intelligence, ability to adapt oneself to different conditions—where such factors as temperament, attitude and interests are of prime importance—must also be considered.

It is now rather common to resort to capacity and ability tests of various kinds in the selection of candidates for different duties. As a result of comprehensive research, various methods have been evolved to obtain reliable measures of a person's ability.

To plot an individual's personality, so-called projective methods have been applied in clinical work, and during recent years they have also been used in the selection of personnel in the economic life of the country. A necessary condition for the use of psychological methods is that their reliability has been tested, and that it is known with a good degree of certainty when and if one can base a judgment on their results. The great need for methods of judging personality has led to the application of projective methods earlier than was really justified by the results of investigations made to ascertain their reliability. This has been the cause of adverse criticism of the reliability of the tests.

It was with this in mind that the Swedish Council for Personnel Administration in 1955 made a grant to the University of Stockholm for studies of the reliability of personality tests based on some of the most common projective tests. One of these investigations, referring to the Thematic Apperception Test (TAT), has been carried out by David Magnusson who, in the following, reports the results achieved.

The Swedish Council for Personnel Administration expresses its thanks to the author and to all others who have contributed to these studies.

Axel Enström
Chairman

Acknowledgements

The study presented here is part of an investigation of the most common projective techniques. The initiative was taken by Professor Gösta Ekman, the Head of the Psychological Laboratory of the University of Stockholm, where the investigation has been performed. I have performed the TAT-study, for which I alone am responsible. Participants in the investigation as a whole have been Fil. kand. Ingrid Brehmer, Fil. kand. Rose Derestad, and Fil. kand. Paul Franér. I thank them for their kind co-operation.

The field work was carried out at five Swedish People's High Schools: Mora, Västerberg, Jönköping, Jära and Tollare. The schools were chosen in collaboration with the Royal Board of Education. The positive attitude of the headmasters and the students of the schools was of great value in this part of the investigation.

During all stages of the investigation the never-failing interest of my teacher and principal, Professor Gösta Ekman has been a valuable support. Professor Kjell Härnqvist has also devoted much time and interest to my problems. I am most grateful to them.

I am also much obliged to Docent Mats Björkman, Docent Jan Gästrin, Fil. lic. Sven-Eric Henricson, Laborator Sten Henrysson, Docent Carl W. Ramfalk, and others too numerous to mention here, with whom I have had the benefit of stimulating discussions during various stages of the investigation.

My sincere thanks are due to the test examiners. Their readiness to perform the ratings made the study possible in its present form.

Fil. kand. Berith Aschberg has assisted me in the statistical work. She has performed her tasks with great care and responsibility. Mr. Albert Read has checked the English. The figures have been plotted by Mr. Tryggve Johansson. I thank them all for their skilful work.

The study has been financed by research grants from the Swedish Council for Personnel Administration and Statens Psykologisk-Pedagogiska institut.

The Psychological Laboratory
The University of Stockholm
April 1959

David Magnusson

Contents

Foreword 3
Acknowledge 5
Contents 7
Tables and appendices 12
Figures 14
Introduction 15

Part I. Criterion data 21

Chapter 1. *Collecting of criterion data* 23
1. Subjects 23
A. Schools 23
B. Groups of students 23
C. Comments 24
2. The students' ratings 25
A. Variables 25
B. Definition of the variables 25
C. Rating scale 27
D. Instruction 27
E. Comments 27
3. Objective test data 28
A. Intelligence test 28
B. Further data 28
4. Field work 28
A. Group situations 29
B. Individual test sessions 29

Chapter 2. *Criterion data* 31
1. The students' ratings 31
A. Transformation of the students' ratings 31
B. Reliability of ratings by students 32
2. Objective test data 33
3. The students' ratings and objective test data 34

Chapter 3. *An analysis of the students' ratings* 35
1. The correlation matrix 35
2. Factor analysis 35
A. Rotations 36
3. Factors 38
A. Factor A 38
a. An investigation of halo in factor A 39
B. Other factors 45
4. The application of the factor analysis in our investigation (TAT) 47
A. Correlation between matrices 47
B. Cluster analysis 49

Part II. Projective test data 51

Chapter 4. *The collecting of the protocols* 53
1. Subjects 53
2. Pictures 53
3. Test sessions 54
A. Instruction 54
B. Recording 55
C. Test length 55

Chapter 5. *Judging of the protocols* 57
1. Test examiners 57
2. Variables 57
3. The test examiners' ratings of the protocols 58
A. Instruction 58
a. Prediction of the students' ratings 58
b. Rating of latent behavior 59
c. Rating of subjective confidence 59
d. Rating of one variable at a time 59
B. Material 60
C. Scoring system 60

Chapter 6. *The test examiners' ratings* 61
1. General remarks 61
2. Transformation of the examiners' ratings 61
3. An analysis of the test examiners' ratings 61
A. Problems 61
B. Calculations 62
C. Cluster analysis 63
a. General discussion 63
b. Definition of cluster 64
c. Clusters 64
d. Comparison of clusters for the test examiners . . 64
D. Agreement between the test examiners on the intercorrelations for one variable at a time 67
a. Procedure and results 67
4. Agreement in structure between the students' and the test examiners' matrices 69
A. General remarks 69
B. Comparison of clusters for examiners and students 70
C. Agreement between examiners and students on the intercorrelations for one variable at a time . . . 71
5. Sources of variance for the ratings by the students and the test examiners 74
A. The variance of the students' ratings 74
B. The variance of the test examiners' ratings 74
6. Ratings of latent behavior. 75
A. General remarks 75
B. The relation between ratings of latent and manifest behavior from the TAT 75
7. Ratings of subjective confidence 77

Part III. Results 79

Chapter 7. *Inter-rater reliability* 81
1. General discussion 81
2. The present investigation 84
A. Statistical procedure 84
B. Results 84
C. Comments 86

Chapter. 8. *Validity* 87
1. Some earlier investigations 87
a. Sanford et.al. 87
b. Hartman 87
c. Harrison 88
d. Davids and Rosenblatt 88
2. The present investigation: Validity of the examiners' ratings with the students' ratings as a criterion . . . 89
A. Statistical procedure and results 89
B. Comments 92
a. Blind analysis 92
b. Differences in content of the variables 92
c. The rating of all subjects in all variables . . . 93
d. Mean validity 93
3. Validity of ratings in the "alienation syndrome" . . . 94
A. Problem 94
B. Procedure and result 95
C. Comments 95
4. The relation between TAT-ratings of latent behavior and the students' ratings 96
A. Problem 96
B. Result 96
5. TAT-ratings and objective test data 97
6. Validity and agreement in structure 98
A. Procedure and results 98
a. Problem I 98
b. Problem II 98
B. Comments 101

Chapter 9. *Reliability and validity* 102
1. General discussion 102
2. A practical consequence 102
A. The present investigation: Procedure and results . . 102
B. Comments 104

Chapter 10. *Rank order between test examiners and variables* 105
1. Rank order between test examiners 105
A. Problem 105
B. Statistical procedure 106
C. Result 106
D. Comments 107
2. Rank order between variables 107
A. Problem 107
B. Procedure and result 108
C. Comments 109

Chapter 11. *Subjective confidence—actual ability* 110
1. General discussion 110
A. Problem 110
B. Some earlier findings 111
2. The present investigation: The relation between subjective confidence and objective accuracy over individuals 111
A. Problem 111
B. Individuals rated with great or little subjective confidence 112
a. Problem 112
b. Procedure and results 112
C. Individuals, rated with great or little objective accuracy 113
a. Problem 113
b. General discussion of procedure 114
c. Procedure and results 114
D. The correlation between subjective confidence and objective accuracy for individuals 116
a. Problem 116
b. Procedure and results 116
3. The present investigation: The relation between subjective confidence and objective accuracy over variables . 117
A. Problem 117
B. Procedure and results 117
4. Conclusions 118

Chapter 12. *Subjective confidence, objective accuracy and test length* . 119
1. Subjective confidence and test length 119
A. General discussion 119
B. Procedure and result 119
C. Comments 120
2. Objective accuracy and test length 121
A. Problem 121
B. Procedure 121
C. Comments 122

Chapter 13. *Test ratings and test length* 123
A. Problem 123
B. Procedure and result 123
C. Comments 125

Chapter 14. *Profiles for individuals rated with high or low confidence and for individuals rated with high and low objective accuracy* 127
1. Profiles for individuals with high and low confidence ratings 127
A. Problem 127
B. Procedure and results 127
C. Comments 129
2. Profiles for individuals judged with high or low objective accuracy 129
A. Problem 129
B. Procedure and result 130
C. Comments 130
3. Test length, ratings of subjective confidence, and test length 131

Chapter 15. *Test length and students' ratings* 132
A. Problem 132
B. Procedure and result 132
C. Comments 132

Chapter 16. *Inter-rater reliability coefficient and validity coefficients with test length partialled out* 134
A. Discussion 134
B. Inter-rater reliability with test length partialled out . 134
C. Validity coefficients with test length partialled out . . 135

Chapter 17. *TAT versus Rorschach and Sentence completion* 137
1. Rorschach 137
A. Ratings, based on the Rorschach protocols 137
B. TAT—Rorschach 138
C. Comments 139
2. Sentence completion 139
A. Variables 139
B. Ratings, based on the SCD 140
C. TAT—SCD 141

Part IV. Summary and suggestions for further research . . 143

Chapter 18. *Summary* 145

Chapter 19. *Suggestions for further research* 150

Appendices and references 153

Appendix A 154
Appendix B 166
Index of abbreviations 172
References 173

Tables and appendices

Tables in the text

Part I

1. Subjcets 24
2. Variables 26
3. Reliability coefficients for the students' ratings 32
4. Means and standard deviations of raw scores on the F test for various groups 33
5. Distribution of residuals after the extraction of six factors 36
6. Rotated matrix (oblique rotation) 37
7. Correlation between primary factors 38
8. Coefficients for the correlation between sociometric rank order and the students' ratings in various variables. Factor loadings in factor A 41
9. Clusters for the students' ratings. Corresponding factors and factor loadings 50

Part II

10. Number of protocols in various groups 53
11. Grouping of clusters for the examiners' ratings I 65
12. Grouping of clusters for the examiners' ratings II 66
13. Grouping of clusters for the examiners' ratings III. 66
14. Agreement between the test examiners as to the rank orders of inter-correlation coefficients in the matrices 68
15. Mean rank order coefficients for the agreement between the test examiners' matrices 69
16. Grouping of clusters for examiners and students 70
17. Agreement between each examiner and the students on the inter-correlations for one variable at a time 71
18. Mean reank order coefficients for agreement in structure between the test examiners and the students and betwen the test examiners mutually 73
19. Correlation between ratings of latent and manifest behavior (Examiner C) 76
20. Number of individuals with different ratings for manifest and latent behavior 76
21. Frequences of double ratings 77

Part III

22. Intraclass coefficients etc. for inter-rater reliability 85
23. Validity coefficients (corrected for attenuation in the criterion) . . 90
24. Validity coefficients for pooled ratings in the alienation syndrome . 95
25. Coefficients for the correlation between the students' ratings and ratings of latent (Cl) and manifest (Cm) behavior for examiner C . 97
26. Coefficients for the correlation between TAT-ratings in variable 6, Acute, and test scores 98

27. Validity coefficients for pooled ratings by two examiners 103
28. Rank orders of validity coefficients for each variable 106
29. Rank orders of validity coefficients for each examiner 108
30. Reliability of pooled ratings of subjective confidence 112
31. Reliability of the measure of objective accuracy for each examiner 115
32. Correlation between subjective confidence and objective accuracy . 116
33. Coefficients for the correlation between ratings of subjective confidence and test length 120
34. Coefficients for the correlation between objective accuracy and test length 122
35. Coefficients for the correlation between ratings by the test examiners and test length 124
36. Coefficients for the correlation between the students' ratings and test length 133
37. Inter-rater reliability coefficients with test length partialled out . . 134
38. Validity coefficients with test length partialled out 135
39. Reliability of pooled ratings based on the Rorschach 138
40. Coefficients for the correlation between TAT-ratings and pooled ratings on the Rorschach 138
41. Inter-rater reliability of pooled ratings based on the SCD (ratings of items) 140
42. Reliability of pooled ratings based on the SCD (over-all ratings) . . 141
43. Coefficients for the correlation between ratings, based on the TAT and on the SCD (ratings of items) 141
44. Coefficients for the correlation between the TAT-ratings and pooled ratings from the SCD (over-all ratings) 142

Tables in the appendix

Appendix A

45. Correlation matrix for the students' ratings I (20 × 20 variables N = 83) 154
46. Correlation matrix for the students' ratings II (16 × 16 variables N = 65) 155
47. Centroid matrix 156
48. Rotated matrix (Orthogonal rotation) 156
49. Transformation matrix (Orthogonal rotation) 157
50. Transformation matrix (Oblique rotation) 157
51. Cosines between reference vectors 157
52. Correlation matrix. Examiner A (N = 63) 158
53. Correlation matrix. Examiner B (N = 63) 159
54. Correlation matrix. Examiner C (N = 63) 160
55. Correlation matrix. Examiner D (N = 63) 161
56. Clusters for examiner A 162
57. Clusters for examiner B 163
58. Clusters for examiner C 164
59. Clusters for examiner D 165

Appendix B

1. An example of the rating forms 166
2. The complete definitions of the variable 166

For the determination of the significance level of values reported in the tables Statistical Tables by Fisher and Yates (1953) has been used.

When nothing else is stated, the levels of significance given in the tables or in the text have been obtained by two tail tests.

Figures

1. Factor loadings in factor A—criterion coefficients 43
2. Factor loadings (factors B, C, D, E)—criterion coefficients 44
3. Plot for the two correlation matrices (16 × 16 and 20 × 20 variables) 48
4. Distribution of validity coefficients for various test examiners . . . 91
5. The relation between agreement in structure for variables and the size of validity coefficients 99–100
6. Distribution of ranks for various test examiners (when the validity coefficients for each variable are ranked in order of size) 105
7. Profiles for individuals judged with high and low subjective confidence 128
8. Profiles for individuals judged with great or little accuracy 130

Introduction

1. GENERAL REMARKS

The Thematic Apperception Test, in the following called the TAT, was constructed as

> "a method of revealing to the trained interpreter some of the dominant drives, emotions, sentiments, complexes and conflits of a personality. Special value resides in its power to expose the underlying inhibited tendencies, which the subject, or patient, is not willing to admit, or cannot admit because he is unconscious of them." (Murray, 1943, p. 1.)

Since the test was published, there has been much discussion of and theorizing on the relationship between the latent tendencies of personality, as they may be expressed in the fantasy behavior in the test protocol and interpreted by the test examiner, and the manifest behavior of the subjects who have told the stories in the protocols. The interest will be traced back to Murray's statement in the manual:

> "... statistical studies have shown that with most variables there is a positive correlation between the strength of their imaginal (TAT) expressions and the strength of their behavioral expressions. We cannot lean very heavily on this over-all finding, however, since not only do we find numerous individual exceptions, but in the case of certain other important drives and emotions, especially those which are customarily repressed, the exact opposite is generally true." (1943, p. 16).

Another of the initial broad explorative investigations in this field presents data with the same tendency as Murray's for the relation between latent tendencies, revealed by the TAT and manifest behavior; a positive relationship for most of the variables, no relationship at all for some variables and a negative one for a few variables (Sanford, 1943, pp. 281–2).

Since the beginning of the use of the test, most of the investigators in the field have stuck to the *lacking* positive relationship of a few variables. Theoretical, experimental and explorative stu-

dies have mostly revolved around this problem for a few variables—especially aggression (see e.g. Wittenborn, 1957, pp. 342–3, who gives a survey of investigations in the field during one year). A number of theories have been advanced, beginning with those by Murray and Sanford. Less interest has been devoted to the positive relationship of most of the variables and to developing methods for interpretation and scoring of the test protocols for these variables in order to use the test for prediction and personnel selection. This circumstance has also led to a growing need for more data about the relationship between manifest behavior and the TAT-responses. Thus Eron (1955, p. 128) concludes in an article:

"The main problem for future research is to determine in a similar way just how fantasy is related to behavior, so that specific predictions can be made from behavior in the controlled testing situation to behavior on the job, in school, at home, with the opposite sex, with like sexed peers, etc."

The use of the test for such purposes is motivated by Bellak (1954, p. 8):

"There are, however, certain situations in which we are legitimately called upon to make inferences from the latent to the manifest: for theoretical, experimental purposes and for such clinical situations as the prediction of criminal behavior and suicide, and such problems as personnel selection."

The need for more data about the possibility of making valid predictions of manifest behavior from the TAT is very urgent, since the technique is not only used in clinical diagnostic work—where such predictions may be of less interest—but also in actual personnel selection where the technique is used in spite of the unsatisfactory validation. This is shown by the following quotation.

"Fortunately, there is a growing number of psychologists, trained in personnel work and experienced in the problems of industry, who see the possibilities inherent in projective techniques as predictive devices. These individuals are making the analysis of projective techniques available in terms that management can understand and use." (Taylor and Nevis, 1957, p. 473.)

One tendency with important consequences for the interpretation of the results from the initial studies in the field is the tendency to try to validate the test as a whole. Thereby positive results for the relationship between latent tendencies as expressed

in the TAT-protocol and manifest behavior were rejected together with the negative results. However, another view on validation of such test instrument is on the way. This view, generally accepted for validation of ability tests for a long time, is expressed by Shneidman (1956, p. 637) in the following statement:

"There is a growing understanding of the complexity of the meaning of validity. As MacFarlane and Tuddenham (MacFarlane and Tuddenham, 1951) have pointed out, there is no answer to such general questions as 'Is the Rorschach valid?' Rather one must be more specific and ask 'Valid with respect to what?' This trend toward specific piecemeal validation is the single most important methodological insight in recent projective test science and merits further and refined exploration. Once the notion was understood that one could not simply validate the Rorschach (or TAT, and so on) but had to validate in terms of specific and more limited criteria, the whole problem immediately became more clear and more manageable."

Shneidman says further on:

"As stated, the attempt to achieve over-all validation of a complex psychological instrument is a chimera. It appears more and more evident that bite-size validation in reference to specific dependent criteria is the more fruitful (albeit less megalomaniac) course."

The relation to current personality theory should be touched upon. Then attention must be drawn to two distressing facts; the lack of one generally accepted theory of the structure and function of personality and the lack of a generally accepted theory of personality and perception. A consequence of these facts is the lack of a common vocabulary with generally accepted definitions of the terms. Meaningful and productive predictions and interpretations of the results are limited by this circumstance. There are many theories claiming to cover and explain the whole personality but each stressing different aspects. The investigators are often bound to one or another of these theories and interpret the results of their studies in the light of a specific theory.

In this situation we think that it is most important to fulfill two objectives: a) to contribute to our knowledge of the specific test instrument and b) to give empirical data from a specific field which may later be used in a general theory of personality and personality assessment, a theory which will take into consideration all the known facts in the field.

2. THE PRESENT INVESTIGATION

A. General characteristics

The main purpose of this study is to investigate the possibility of making predictions of ratings, based on observations of manifest behavior from the fantasy behavior in the TAT protocols. With the earlier discussion as a background the investigation may be characterized as dealing with validity. We want to contribute to the knowledge of one aspect of the test. We have studied the relationship between ratings of male subjects for a number of variables—with the ratings performed by fellow students on the basis of two years' acquaintanceship—and ratings obtained by blind analysis of TAT protocols as predictions of the students' ratings. The students' ratings then serve as a criterion of the validity of the ratings based on the TAT. The validity studied is thus restricted to the aspects defined by the character of the criterion. The characteristics of the criterion will define the validity studied in this way as *concurrent* validity (Technical recommendations 1954, p. 14). The relationship between the ratings based on the TAT and ratings based on some other projective methods is calculated. Some irrelevant factors affecting the rating procedure are investigated. These calculations may contribute to the knowledge of the *construct* validity of the method (Cronbach and Meehl, 1955, pp. 287–8, Härnqvist, 1959, pp. 11–13). Some problems related to the main ones are studied.

The investigation is explorative in character. No hypotheses formulated before the beginning of the investigation have been tested.

The study of the TAT is part of an investigation, in which the same main criterion has been used for the study of the same aspects for four other projective methods: the Rorschach, the Sentence completion test, a colour test, Pfister's Der Farbpyramidentest (Heiss, R. und Hiltmann, H., 1951) and a small projective test, the ring-test, constructed by Lindberg (1947, 1949). This fact has had some consequences for the design of the TAT investigation. Separate accounts have been delivered for the Sentence completion test and the colour test (Derestad, 1958, and

Brehmer, 1959). The account of the investigation with the Rorschach is in preparation.

Through its exploratory character the investigation will touch upon many problems concerning the reliability and validity of the TAT method. These problems have been discussed and investigated from various aspects earlier. Only such investigations and discussions which bear directly on the aspects studied are referred to in the account of the present investigation.

B. Plan of the report

We have chosen to account for the present investigation according to the following general plan:

PART I contains an account of the procedure of collecting the ratings by the students and an analysis of these ratings.

PART II concerns the test ratings, and follows the same general plan as Part I; collection and analysis of the ratings.

PART III gives an account of the results found for the problems studied; inter-rater reliability, concurrent validity, the role of the test length in the rating procedure, subjective confidence and objective accuracy. Part III ends with a short account of the relation between the TAT-ratings and the ratings based on the Rorschach and the Sentence completion test.[1])

PART IV summarizes procedures and results. In the last chapter attention is drawn to possible future research, to which these results may be applied.

[1]) The data based on the Rorschach and the sentence completion test have been used with kind permission from the administrators of these tests, Fil. kand. Paul Franér and Fil. kand. Rose Derestad.

PART I

Criterion data

Chapter 1. Collecting of criterion data

1. SUBJECTS

The subjects in the investigation were sixty-four male students from second-year groups at four Swedish People's High Schools (folkhögskolor).

A. Schools

The schools at which the investigation took place are a type of school confined almost solely to the Scandinavian countries (see The Royal Board of Education, 1958, pp. 51–54). They are boarding schools open to both male and female students and with two- or three-year courses. The schools are voluntary. The mean age of the students at the beginning of the first year is about twenty-one years (cf. Folkhögskolan och de stora årskullarna, 1956, p. 16). Most of the students have not attended any school, after finishing their basic compulsory education at fourteen years of age, and have been employed in practical work for a number of years. Some of the students have already decided what vocation they will follow. Others have decided to change their occupation but have no definite ideas about what they want to do. They often hope to arrive at a decision during the course at the school.

Various investigations indicate that the mean intellectual level of the students at these schools is somewhat higher than that of the total national population. It may be assumed, too, that the range of intelligence is less wide than in the total population. This is presumably also the case, as regards the social standing of the parents of the students. The middle class is probably somewhat over-represented (Jonsson, 1957, pp. 41–42).

B. Groups of students

On account of the boarding house character of the schools, the way of working, and the social life, the students get to know each

other fairly intimately. Advantage has been taken of this circumstance in the investigation. Every student in each group has been rated by all the other students of the same group for all the twenty variables, which were constructed. Male students were only rated by male students. At certain schools the number of ratings that each student had to make, if all male students were to rate all other male students, was too great to be made within a reasonable time. Therefore, at one school where the number of students exceeded twenty, the individuals were divided into two groups according to the intimacy of the relationships between the individuals. Thus the ratings were made in five groups at four schools.

The distribution of students in various groups is shown in Table 1. The numbers given represent those who participated in any rating or testing. The table also shows the mean age of the members of the groups and the total.

TABLE 1. *Subjects*

Group	I	II	III	IV	V	Σ	M
Number	13	18	10	13	10	64	12.6
Age	21.8	21.8	22.7	21.5	23.7		22.5

As is seen, the mean age is about the same as that which is usual for second-year groups (p. 23).

C. Comments

By choosing students for the behavior ratings who have known each other for two years, we have fulfilled the four of Cattell's methodological points for good rating, that are relevant to the problems dealt with in this chapter, namely, implicit time sampling, time length, avoidance of role relations, and democratic judgment (Cattell, 1957, pp. 63–69). The first point claims that

"the subject should live with the rater (as in a fraternity) or be under observation for so much of the day that the former is seen in many roles and many stimulus conditions, as required in actual time sampling."

The point of time length means that

"the period of observation should extend to at least two or three months

(preferably a year) so that a sufficiency of extreme situations are encountered to get at deeper emotional reactions."

The best way of avoiding special role relations is, according to Cattell, to use peers as raters of behavior. The rating procedure ought to be "democratic"; therefore it should be performed by ten to twenty judges, whose values are averaged.

2. THE STUDENTS' RATINGS

A. Variables

The variables, for which the validity of the test is investigated, were constructed according to three principles. First, it should be possible to rate them from the test protocols, second, it should be possible for the students to rate them and third, they should cover the personality of the subjects as closely as possible. The following procedure was followed.

From the literature on TAT, Rorschach and the colour test, a number of variables were collected. Clinical psychologists with long experience of the tests judged these variables with regard to their rateability when the TAT, Rorschach and the colour test are applied. They expressed their judgment on which of the variables presented were a) easy to rate, b) possible to rate and c) impossible to rate from the tests. On the basis of these findings the final set of variables for the investigation was put together during a number of discussions.

For the Sentence completion test the items were constructed after the establishment of the variables for the other tests, and with the definitions of the variables as a basis.

B. Definition of the variables

Each variable was first defined by a general definition, which was made as unambiguous as possible. The definitions were constructed so as to assign as little inference and interpretation as possible to the raters, without identifying the variables with single items of behavior.

A seven-point scale was used for the ratings. Standard-forming definitions were given for the highest and the lowest points of the scale. One definition is given as an example.

Variable 6

Logical reasoning

Ability to draw conclusions and reason logically varies greatly. Some people think clearly and logically, combine rapidly, and are able to advance their views correctly and lucidly.

Some people, however, find it difficult to think coherently and consistently, and cannot easily arrive at valid conclusions.

Most of us are average in this respect.

1. Dull	4. Average	7. Acute
Thinks unclearly		Draws valid conclusions
Combines badly		Combines well

The complete definitions of the variables are given in Appendix B (p. 164). In Table 2 are given defining expressions for both directions of each variable. When nothing else is stated the definitions of the scale value 7 are used in the following text.

TABLE 2. *Variables*

	Scores	
Variable	1	7
1.	Intropunitive	Extrapunitive
2.	Reticent	Ostentatious
3.	Submissive	Dominant
4.	Believing in authority	Oppositional
5.	Cautious	Impulsive
6.	Dull	Acute
7.	Well-adjusted	Maladjusted
8.	Negligent	Careful
9.	Unscrupulous	Scrupulous
10.	Insensitive	Sensitive
11.	Dependent	Independent
12.	Gloomy	Cheerful
13.	Objective	Subjective
14.	Confident	Anxious
15.	Unsociable	Sociable
16.	Not reflective	Reflective
17.	Cold, Indifferent	Warmhearted
18.	Lacking vitality	Vital
19.	Tolerant	Intolerant
20.	Rational	Emotional

As seen from Table 2 the definitions for the scale value 7 for some variables represent the direction of the variables that might be considered as the positively value loaded direction, and for some the direction that might be considered as the negatively value loaded direction.

C. Rating scale

The ratings were made on a seven-point scale. Forms of the type shown in Appendix B (p. 164) were used. The order of the variables was randomized between the students.

D. Instruction

Each student rated all subjects of the group in one variable before going on to the next. The students were instructed first to mark the individuals rated as the most extreme in both directions with the scale value of 1 or 7, then to mark those subjects who, after that, were the most extreme in both directions on 2 and 6, and so on. Those who did not deviate in any direction should finally be marked on the scale value 4. The students were not compelled to give a definite number of ratings in each category but were told to aim at a normal distribution of the ratings for each variable. If an extreme rating on any variable should appear absurd to the rater, it would not be necessary to mark that extreme. For training the instruction included the rating of a variable (the ability to recite poetry).

E. Comments

The performing of the ratings in one variable at a time for all individuals in the group—by which attention is more directed to the behavior variable rated than to the individual—was carried out in order to reduce possible stereotype effects (Cattell, 1957, p. 65). The same purpose was fulfilled by the changing of direction for the "positive" definitions of the variables. In an experimental study of the "halo" effect, Johnson and Vidulich have shown, with empirical data, that the rating of one variable at a time over all individuals, instead of rating one individual at a

time on all variables, will have the effect of reducing the proportion of the variance, that may be referred to as "halo" (Johnson and Vidulich, 1956, pp. 130–4).

3. OBJECTIVE TEST DATA

A. Intelligence test

In order to gauge the intellectual level of the students, a group intelligence test was administered. The F-test, constructed by a team under Härnqvist, was used (Härnqvist, 1956). This test has been standardized on students from the same type of schools as those in the present investigation. As only a limited time was available, only two of the sub-tests were administered. They were chosen so as to cover the main aspects of the definition of variable 6, Acute (p. 26). The first, a synonym test of multiple choice type, is assumed to measure the verbal factor. Its reliability was estimated at .90 in a split-half coefficient at the standardization. The other test is a test for measuring the reasoning factor of intelligence. The reliability of this test is given in a split-half reliability of .85.

B. Further data

In order to get information about the students' personal circumstances a questionnaire was constructed in which information was obtained about the subject's age, the age and profession of the parents, number of siblings, earlier schools and so on.

(A Swedish translation of the Cornell Index was also administered (Weider et.al., 1949). Since the Swedish translation has not been standardized in any way, the results are not given here.)

4. FIELD WORK

The collecting of the data was divided into two stages; first, collecting all criterion data and the Sentence completion test in group situations, and second, testing with the other projective methods in individual sessions.

A. Group situations

One whole day was assigned to the group sessions at each school. The day began with a general orientation for the students about the purpose of the investigation. The experimenter pointed out that the ratings and the results would only be used in the investigation and under safe control, so that no unauthorized person would have access to the material, nor would it be used for other purposes.

In order to comply with the demand that no unauthorized person could get hold of the material and identify the results and ratings, all the forms used were marked with a code number for each individual. All the forms for each student were handed in in an envelope, which was also marked with the code number.

All group test sessions were administered by the same test leader. Every session followed the same plan:

1. Orientation
2. Sentence completion (about 35 minutes)
   Pause
3. F-test (30 minutes)
   Pause
4. Questionnaire (35 minutes)
5. Cornell Index (10 minutes)
   Lunch
6. Rating by Subjects.

B. Individual test sessions

The individual test sessions began at the first schools immediately after the administration of the group sessions. The Rorschach test and the colour test were given by one tester, while the TAT and the ring test was administered by another. In two groups all the individuals were tested with the Rorschach and the colour test before the other tests were administered. In two other groups the order was reversed. In the fifth group all the individual test sessions were given during the same period of time.

A detailed report of the TAT-testing is given in connection with the account of the test data.

Here it may be pointed out that on the whole both teachers and students at the schools were very positively inclined towards the investigation. The results of the investigation must be judged in the light of the fact that the protocols are obtained from well-motivated subjects.

Chapter 2. Criterion data

1. THE STUDENTS' RATINGS

A mean of the obtained ratings was computed for each individual on each variable. This value was used as an estimate of the extent to which the individual possessed the trait in relation to the group standard (see Cattell, 1957, p. 64). The means and standard deviations of these values were computed for each variable and each group.

A. Transformation of the students' ratings

The students' ratings were performed with the group standard as the norm (p. 27). As the instruction did not claim a forced distribution, there will be some differences between the means of various groups. Since the groups have not been compared with each other on a common scale, the differences obtained cannot be expected to reflect possible true, predictable differences between the means of the various groups. Therefore the variance of the differences between the group means ought to be excluded from the calculations based on the students' ratings.

The reasons given above will lead to the conclusion that the best way of computing correlation coefficients based on the students' ratings is to calculate one coefficient for each group. One mean coefficient for each variable could then be obtained by transformation to Fisher's z (Edwards, 1950, pp. 133–4). In some calculations this procedure has been followed.

On account of the great number of coefficients to be computed, some calculations have been performed on an electronic data machine. Then we wanted to compute the correlation coefficients over all individuals at a time on each variable. Therefore a total distribution was formed of the standard scores in each group. This total distribution was transformed to a distribution with 15.5 as the mean and 5.0 as the standard deviation in order to be used in

the data machine. The variance due to differences between the means for various groups is excluded from the correlation coefficients. The coefficients, calculated on the transformed ratings over all individuals at a time, thus give the correlation within groups. The degrees of freedom = N — m — 1, where m is the number of groups (Lindquist, 1940, pp. 221–2).

The exclusion of the variance between groups from the calculations does not mean that no true differences exist between the group means. As long as we do not know the direction or the size of possible differences, however, the best thing is to give all the groups the same mean and the same standard deviation. As far as such true differences do exist, the coefficients based on the transformed ratings and calculated over all individuals at a time may be too low. Correctly the variance between groups in such a case should have been included in the computation, and the effect of its inclusion should have been to extend the correlation surface and increase the size of the coefficents.

B. Reliability of ratings by students

The reliability of the students' ratings was estimated by split-half technique. The coefficients were computed for one group at a time on each variable. One mean coefficient for each variable was calculated by transforming the coefficients to Fisher's z. In these coefficients the irrelevant variance, due to differences between the obtained means of the groups, is not included. The reliability coefficients, obtained by Spearman-Brown's formula, are given in Table 3.

TABLE 3. *Reliability coefficients for the students' rating*

1. Extrapunitive	.82
2. Ostentatious	.92
3. Dominant	.92
4. Oppositional	.81
5. Impulsive	.87
6. Acute	.87
7. Maladjusted	.58
8. Careful	.89
9. Scrupulous	.89
10. Sensitive	.85

11. Independent	.88
12. Cheerful	.86
13. Subjective	.81
14. Anxious	.76
15. Sociable	.92
16. Reflective	.91
17. Warmhearted	.67
18. Vital	.81
19. Intolerant	.68
20. Emotional	.71

The median reliability coefficient is .855. On the whole we may conclude that the coefficients are of sufficient size to admit the use of the ratings as estimates of the positions of the individuals on the continua. The smallest of the coefficients is that for variable 7, Maladjusted, which is .58. This value will give a standard error of measurement of .648 SD.

2. OBJECTIVE TEST DATA

The verbal and the reasoning factors were measured by the synonym and the reasoning tests as mentioned above. The means and the standard deviations for the five groups are given in Table 4. There, too, are given for comparison the means for the standardization sample. As is seen from the table, the means for all the groups but one are higher on both the sub-tests, than those for the standardization sample.

TABLE 4. *Means and standard deviations of raw scores on the F test for various groups*

Group		I	II	III	IV	V	M	Stand. sample
Subjects		12	16	10	13	10		
Synonyms	M	24.9	28.4	25.9	28.6	26.9	27.1	23.8
	SD	5.55	3.35	5.45	5.17	3.37	4.87	
Reasoning	M	15.8	17.1	14.2	15.3	15.6	15.2	14.6
	SD	2.50	3.27	4.28	5.08	3.17	3.90	

This is what might be expected, since the standardization was performed on students in the first-year group at People's High

Schools. A certain selection, based partly on the intellectual level of the students, takes place from the first to the second year. The school training may also influence ability to perform the sub-tests, especially the synonym test.

3. THE STUDENTS' RATINGS AND OBJECTIVE TEST DATA

In the group test data we obtained an objective measure of the intellectual level of the students. The two sub-tests, synonyms and the reasoning, may be considered as measures of two important aspects of the function defined in variable 6, Acute. The correlation between the students' ratings in this variable and the test scores has therefore been estimated.

The students' ratings of each other were performed in relation to the group standard, while the result of the group test gives the positions of the individuals in relation to a general standard. Therefore the coefficients for the correlation between the students' ratings in variable 6 and the test scores were computed for one group at a time. The calculation was based on the pooled standard scores from the two sub-tests. The coefficients obtained were transformed to one coefficient by Fisher's z. This coefficient (.46) is significant at the 1 per cent level.

The size of the coefficient indicates that the students' ratings in variable 6 have good validity, especially in the light of the fact that the students' task was not to predict the test scores but to rate each other according to the definition of variable 6, Acute.

Chapter 3. An analysis of the students' ratings

The main criterion on the validity of the test examiners' ratings consists of the students' ratings of each other. Important conditions for the use of the students' ratings as the criterion are the coherence and consequence, with which the ratings have been performed. A meaningful structure in the correlation matrix of the students' ratings will be the result if these conditions are fulfilled. This chapter is devoted to an analysis of the structure.

1. THE CORRELATION MATRIX

The ratings in each variable were correlated with the ratings in all other variables. The correlation matrix is given in Table 45 in Appendix A (p. 152).[1])

The coefficients in the matrix of Table 45 were computed on the raw scores over all individuals in each variable, whereby the variance due to differences between the obtained means of various groups is included in the coefficients. (The analysis was performed before the transformations described on page 31 were carried out. This is the reason for the computational procedure followed here.)

2. FACTOR ANALYSIS

For a thorough analysis of the structure of the students' ratings a factor analysis of the correlation matrix for the ratings was performed.

The factor analysis was performed according to the complete

[1]) The coefficients in this matrix are based on 83 individuals from six groups of students. This was the number of subjects in the investigation from the beginning. The original plan for the calculations could not, however, be fulfilled, since some circumstances, which might seriously affect the results for this group, were revealed at one school at the time of the TAT investigation. When we decided to change the calculation plans, the factor analysis had already been performed.

centroid method (Thurstone, 1947, pp. 161–70). By this method six factors were extracted. The factors of the centroid matrix are given in Table 47 in Appendix A (p. 154).

In order to decide when a sufficient number of factors had been extracted two criteria were used. The first, Tucker's Phi-coefficient (Fruchter, 1954, pp. 77–79), is based on the degree of change in the sum of residuals after each factor extracted. The other criterion, proposed by Holzinger and Harman (Holzinger & Harman, 1941, p. 22) is based on the relation between the standard deviation of the residuals and the number of individuals, on which the correlation coefficients in the matrix were computed. According to Holzinger and Harman's criterion the extraction of factors could have been finished after the extraction of the fifth factor. As Tucker's criterion did not admit this, a sixth factor was extracted.

The residual values after the extraction of six factors are given in Table 5. The standard deviation of the residuals is .030.

TABLE 5. *Distribution of residuals after the extraction of six factors*

Residuals	Number
−.06	18
−.05	14
−.04	14
−.03	34
−.02	37
−.01	44
.00	70
.01	56
.02	32
.03	27
.04	26
.05	14
.06	6
.07	8
	Total 400

A. Rotations

The centroid matrix was first orthogonally rotated. The factor loadings from the orthogonal rotation are reported in Table 48

in Appendix A (p. 154). In Table 49 in Appendix A (p. 154) is given the transformation matrix for the orthogonal rotation.

The plotting of the factors in the rotated matrix revealed that the factors were obviously correlated with each other to some extent. For this reason an oblique rotation of the matrix was performed.

The oblique rotation was performed according to the single plane method (Thurstone, 1947, pp. 216–24). The rotation was performed in accordance with the criteria proposed by Thurstone (1947, p. 335).

The transformation matrix for the oblique rotation is given in Table 50 in Appendix A (p. 155).

In Table 6 is reported the factor matrix after the oblique rotation. The factor loadings are given in two decimals. The transformation, however, was performed in three decimals.

TABLE 6. *Rotated matrix (Oblique rotation)*

	A_1	B_1	C_1	D_1	E_1	F_1	h^2
1.	–.10	.49	–.02	.01	–.03	–.03	.26
2.	.19	.29	.42	–.02	–.14	.18	.35
3.	.40	.46	.23	–.02	–.03	.07	.43
4.	.51	.50	–.01	.18	–.19	–.17	.62
5.	–.08	.01	.41	–.01	–.43	.14	.38
6.	.74	.33	–.02	–.08	.12	–.01	.57
7.	.06	–.02	–.57	.45	–.13	–.01	.55
8.	–.02	.03	.11	.15	.63	–.10	.44
9.	–.05	–.20	–.20	.00	.53	.14	.38
10.	.39	.10	.02	.44	.23	–.14	.43
11.	.71	.42	–.16	.07	.11	–.02	.73
12.	–.04	.00	.90	.05	.17	–.16	.86
13.	–.56	.00	–.08	.21	.03	.07	.37
14.	–.64	–.23	–.27	.32	.06	–.09	.64
15.	.04	.00	.88	–.01	.00	.15	.80
16.	.54	.00	–.46	.05	.05	.16	.54
17.	–.04	–.50	.24	.19	.25	.18	.44
18.	.49	.06	.07	–.16	.14	.18	.32
19.	–.43	.23	–.11	–.02	.01	.12	.26
20.	–.08	–.33	.00	.64	–.05	.02	.53

As is seen from the factor matrix the loading with the highest value is .18 in factor F_1. This factor may be accounted for by error variance and is thus excluded from further discussion of the factors.

Cosines between the reference vectors are reported in Table 51 in Appendix A (p. 155).

TABLE 7. *Correlation between primary factors*

	T_{A_1}	T_{B_1}	T_{C_1}	T_{D_1}	T_{E_1}	T_{F_1}
T_{A_1}	1.00	–.50	–.10	–.19	.59	–.10
T_{B_1}		1.00	.22	.64	.49	.63
T_{C_1}			1.00	.19	–.21	.27
T_{D_1}				1.00	–.37	.73
T_{E_1}					1.00	–.10
T_{F_1}						1.00

Table 7, where the coefficients for the correlation between the primary factors are reported, shows that some of the factors are relatively highly correlated with each other. Of the factors the purest in this respect is factor C, whose highest correlation is .22 with any of the other four factors, when factor F is not considered. The coefficients are, however, not so large as to invalidate the use of the factors.

3. FACTORS

It is always difficult to interpret factors resulting from factor analysis. This will especially be the case when the analysis is based on ratings. Therefore the discussion of the factors in this case is restricted to one factor. The other factors are only described.

The direction of the positive factor loadings are given in the account of the factors.

A. Factor A

This factor has its highest loadings in the following variables.

Variable	Factor loadings
6. Acute	.74
11. Independent	.71
14. Confident	.64
13. Objective	.56
16. Reflective	.55
4. Oppositional	.51
18. Vital	.49
19. Tolerant	.43
3. Dominant	.40
10. Sensitive	.39

The next factor loading is .19 for variable 2, Ostentatious.

a. *An investigation of halo in factor A.* In most rating procedures, where the variables of the judged objects may be regarded as value loaded, it is usually assumed that the size of the correlation between the ratings on different variables is influenced by the halo effect. This means that behind the ratings in two or more variables lies a common valuation, which affects the ratings. The ratings of any variable are led in the direction of the general impression of the individuals. Thereby a spurious amount of positive correlation between the variables is introduced. There is reason to assume that the more value loaded the judged variables are, the greater will be the influence of the halo factor. The valuation, common for certain variables, will give the intercorrelations an error component which will give erroneously high coefficients. The coefficients are partly dependent upon something other than the objective correlation between the variables. Part of the variance in the correlation matrix for the relation between the students' ratings of each other may thus be referred to the halo factor. This has been discussed by Ekman (1946, pp. 58–65). He concludes, on the basis of empirical findings in a factor analysis of subjective ratings, that one factor may be explained as the result of the error contribution from the halo factor. He calls the factor a fictitious factor *(skenfaktor).*

Recently the halo effect has been treated by Johnson and Vidulich (Johnson and Vidulich, 1956, pp. 130–4) in a well-controlled study on experimental manipulation of the halo effect in ratings.

In summing up the discussions about the halo effect from the time it was first noticed by Wells, 1907, and named by Thorndike, 1920, up to the time of the study, the authors conclude that no convincing proof of the existence of halo has been given. Their results show however, that halo is in part a phenomen of judgment and that the common recommendation for reducing it by changing the conditions of judging is effective.

With the results of Johnson and Vidulich as a background, Cattell states that there is no evidence of any generation of a stereotype factor (Cattell, 1954, p. 65). This is, however, to draw too far-reaching conclusions on the basis of the results. The variance for the interaction between raters and ratees, which is used as an indicator of the halo effect, is certainly not significant at the five per cent level, but is in close proximity when all proper precautions are taken to prevent the halo factor influencing the ratings. The Johnson and Vidulich results show, a) that without careful precautions there occurs a stereotype, which is what we usually call halo, and b) that the effect diminishes with proper precautions, in this case by performing the ratings for all individuals at one variable at a time, instead of rating one individual at a time on all variables. What is not shown, is that the effect is zero under these last conditions.

As the students' ratings are our main criterion of the test ratings, we have found it relevant to investigate the possible role of halo in the ratings.

We have made use of a technique reminiscent of a method proposed by Eysenck (1950, pp. 38–53), and called by him criterion analysis. Eysenck proposes that in verifying a hypothesis about the interpretation of a factor in a factor analysis, we may use a criterion column, consisting of correlation coefficients that express the discriminating power of each variable in accordance to the factor. The coefficients of the criterion column are found by correlating the scores in each variable with an outside criterion of the factor hypothesized.

Simultaneously with the collecting of the projective test data, some sociometric data were collected. Among other things the students judged their fellows as leaders of the group, in a situation where no specific variables should give special merit for a rating.

In four of the groups the judgments were given through the method of paired comparisons. In the other two groups the students rank ordered each other. The split-half reliability was calculated for each group. A mean reliability coefficient for the ratings was computed by means of Fisher's z. This coefficient is .93.

Now, if there exists a common valuation behind the ratings of each student in the personality variables by his fellow students, it is plausible to assume that this same valuation lies behind and influences the rank order resulting from the sociometric ratings. In accordance with the discussion by Ekman we hypothesize that the variance of the first factor may to some degree be referred to halo. Then we will get the greatest agreement between the students' ratings and the sociometric rank order for those variables with high loadings in the first factor, if we correlate the sociometric rank order with the students' ratings in one variable at a time. The agreement, as expressed in the correlation coefficients, ought to be proportional to the size of the factor loadings for factor A in the various variables. This proportionality may be tested by plotting and correlation technique.

The coefficients for the correlation between the students' ratings in each variable and the sociometric rank order were computed for one group at a time with rank-order coefficients. The coefficients were transformed to one coefficient for each variable by means of Fisher's z. The coefficients are reported in Table 8.

The variables are designated with the name of the direction that gives positive correlation with the sociometric status. As seen, the coefficients are positive for the direction of the variables, which are usually considered as highly valuable. This supports our use of the sociometric ratings as an expression of the valuation of the individuals.

TABLE 8. *Coefficients for the correlation between sociometric rank order and the students' ratings in various variables. Factor loadings in factor A.*

Variable	r	Factor loadings
1. Intropunitive	.17	.10
2. Ostentatious	.29	.19
3. Dominant	.43	.40

4. Oppositional	.20	.51
5. Cautious	.17	.08
6. Acute	.70	.74
7. Well-adjusted	.33	−.06
8. Careful	.37	−.02
9. Scrupulous	.24	−.05
10. Sensitive	.51	.39
11. Independent	.64	.71
12. Cheerful	.41	−.04
13. Objective	.57	.56
14. Confident	.70	.64
15. Sociable	.39	.04
16. Reflective	.34	.54
17. Warmhearted	.37	−.04
18. Vital	.61	.49
19. Tolerant	.35	.43
20. Rational	.10	.08

In the last column are given the factor loadings in the various variables of factor A. The factor loadings are reported for the direction of the variables given in the designation of the variables.

To facilitate interpretation we give the highest factor loadings in factor A in order of size, and the corresponding coefficients for the correlation between sociometric status and the students' ratings.

Variable	Factor loadings (A)	r
6. Acute	.74	.70
11. Independent	.71	.64
14. Confident	.64	.70
13. Objective	.56	.57
16. Reflective	.54	.34
4. Oppositional	.51	.20
18. Vital	.49	.61
19. Tolerant	.43	.35
3. Dominant	.40	.43
10. Sensitive	.39	.51

As is seen from the table, the variables that have factor loadings in factor A of any size of importance also show, with one excep-

tion, significant coefficients for the correlation between sociometric status and the students' ratings.

On the other hand there are some variables that have relatively high criterion coefficients but have insignificant factor loadings. They are seen along the y-axes of the plot in Figure 1, where the factor loadings in factor A and the criterion coefficients are plotted. The low factor loadings for these variables may be explained as a result of the principles of simple structure, followed in the rotation.

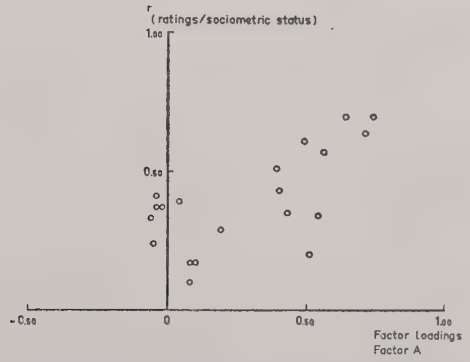


Fig. 1. Factor loadings (factor A)—criterion coefficients.

Already from Table 8 and Figure 1 we can see that the degree of correspondence between the variables plotted is rather good. In order to get a numerical value for the agreement, we computed a coefficient of correlation between the values in Table 8, in accordance with the method used by Eysenck. The agreement is given in a coefficient of .67.

For control the coefficients for the correlation between sociometric status and students' ratings, given in Table 8, were plotted against and correlated with the factor loadings of factor B, C, D and E. The plots are given in Figure 2. The coefficients were .26, .12, .06 and −.11 for factor B, C, D and E, respectively. As is seen from the plots, there is no tendency to agreement except for factor A. The coefficients are all low and of insignificant size.

It may be pointed out that the rotation of the factors was performed quite independently of the criterion coefficients. Eysenck

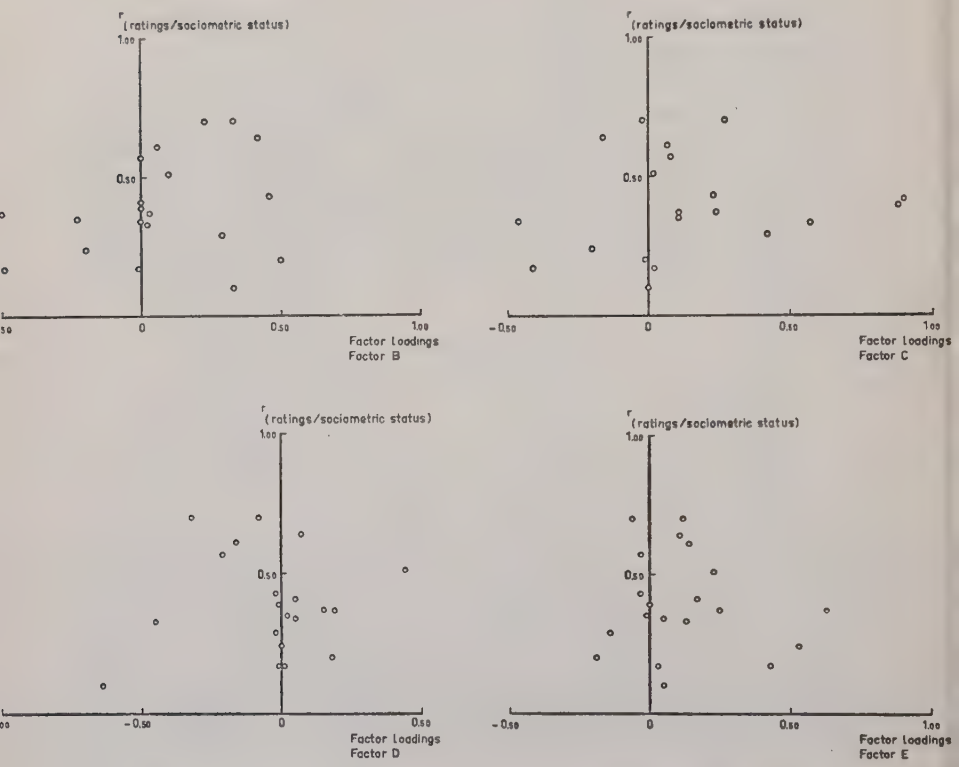


Fig. 2. Factor loadings (factors B, C, D, E)—criterion coefficients.

proposes in the discussion of the method of criterion analysis that the rotation shall be performed so as to give the highest possible agreement between the factor loadings and the coefficients used as criteria (Eysenck, 1950, p. 43). The correlation between the rotated factor loadings and the criterion coefficients was, in his case, .574, when the rotation was performed in accordance with the principle stated. The factor analysis was performed by Eysenck on a 16×16 matrix. If we had rotated according to the principle stated by Eysenck, the agreement would reasonably have been even better than that obtained.

In our case, the sociometric ratings are used as a criterion of the degree of value orientation for each variable. Evidently the sociometric ratings are not merely the results of valuation of the indi-

viduals without any objective basis. The valuation of the individuals in the sociometric ratings follows most probably from the greater existence and coherence of certain highly valuable variables in some individuals—who are thus judged high in sociometric status—than in others, who are judged low. There is a true, objective basis for the valuation. However, with the common instruction for the sociometric ratings, without any specific variable stressed, we have maximized the possibilities of the functioning of the halo factor. Therefore we may expect—above the objective degree of occurrence of the variables in the individuals—a common valuation, which to some degree determines the variance of the sociometric rank orders.

The results support our assumption that the variance, which gives rise to factor A, is—in addition to the actual, objective degree of agreement in the occurrence of the variables in the factor—to some degree based on variance that may be referred to as halo. This will diminish the use of the students' ratings as criterion on the ratings performed by the test examiners. It is not to be expected that the test examiners are able to predict the part of the variance due to halo. The validity coefficients must be interpreted in the light of this fact. However, the results do not invalidate the students' ratings as criteria. As Johnson and Vidulich have shown, the halo effect is minimized by the precautions we have taken. The factor analysis also shows that there exists a fairly good, meaningful structure in the correlation matrix, with four factors besides factor A.

B. Other factors

Factor B. This factor has its highest loadings in the following variables:

Variable	Factor loadings
4. Oppositional	.50
17. Cold, Indifferent	.50
1. Extrapunitive	.49
3. Dominant	.46
11. Independent	.42

Most characteristic of the given direction of the variables is the lack of warm feeling and contact with the environment. Especially those four variables which are most highly loaded with the factor are stamped by the attitude "against" the environment.

Factor C. This factor is the "purest" of the factors. Its highest correlation is .22 with factor B.

Variable	Factor loadings
12. Cheerful	.90
15. Sociable	.88
7. Well-adjusted	.57
16. Not reflective	.46
2. Ostentatious	.42
5. Impulsive	.41

The variables in factor C are typical of a positive, optimistic attitude towards life and others.

Factor D.

Variable	Factor loadings
20. Emotional	.64
7. Maladjusted	.45
10. Sensitive	.44
14. Anxious	.32

The next factor loading in size is that for variable 13, Subjective (.21). Factor D is highly correlated with factor B (.64). The variables indicate personality problems.

Factor E. Factor E has its highest loadings in the following variables:

Variable	Factor loadings
8. Careful	.63
9. Scrupulous	.53
5. Cautious	.43

The next variable in order of size is 17, Warmhearted (.25). The factor correlates .59 with factor A and .49 with factor B.

The directions given may be interpreted as indicating compulsive-obsessional tendencies.

As a whole, we may conclude that our analysis has shown that there exists a meaningful structure in the ratings made by the students.

4. THE APPLICATION OF THE FACTOR ANALYSIS IN OUR INVESTIGATION (TAT).

The factor analysis accounted for in the preceding section was performed for a correlation matrix with twenty variables, where the coefficients were based on ratings of eighty-three subjects (see p. 35). The correlation coefficients in the matrix were computed over all subjects on the raw scores, with no correction for differences in means and standard deviations of various groups. However, in the TAT investigation we have used the ratings on sixteen variables for 65 individuals as our criterion (see p. 35 and p. 58). These ratings have been transformed to a common scale with the same mean and standard deviation for the groups (p. 31). The intercorrelation coefficients based on the transformed scores have been computed and are given in the matrix in Table 46 in Appendix A (p. 153).

Now the problem is: How applicable is the result of the factor analysis of the first matrix on the correlation matrix with sixteen variables, and with the correlation coefficients computed on transformed values for sixty-five individuals?

A. Correlation between matrices

In order to determine how well the correlation matrix on the sixty-five individuals agreed with the primary matrix, two methods were applied. The intercorrelation coefficients for the sixteen common variables from the two matrices were plotted against each other and a coefficient for the correlation was computed. A cluster analysis was also performed.

The plot is given in Fig. 3. As seen from the plot, the coefficients in the second matrix are generally lower than those from the original matrix.

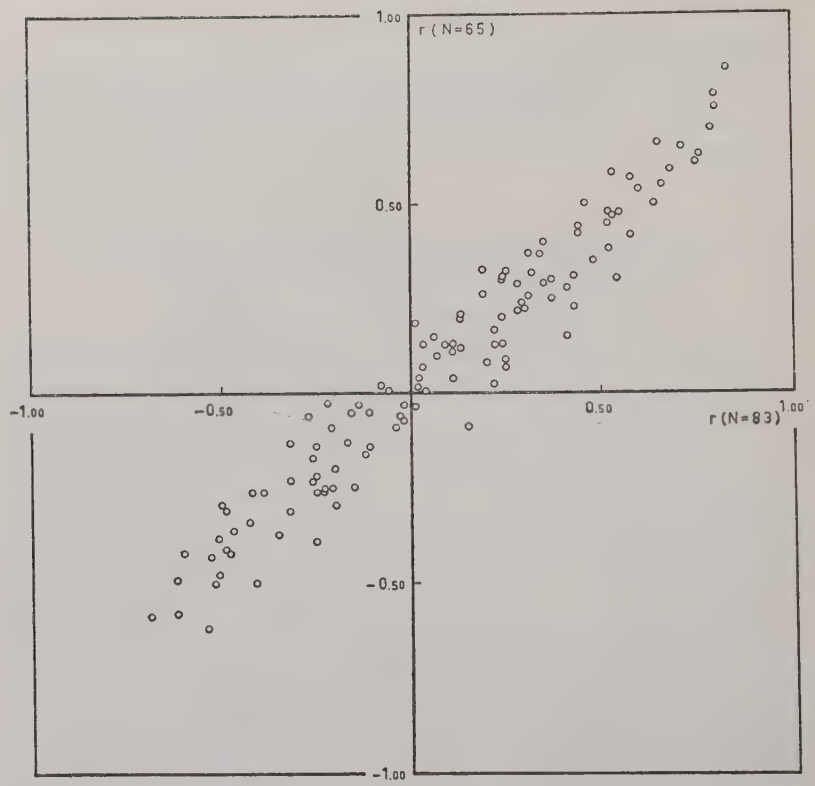


Fig. 3

From the beginning of the investigation, the variables were given their direction at random. Another constellation of directions for the variables should have given other signs to the coefficients in the matrix without necessarily affecting the size of the coefficients. Change of signs for the coefficients in the matrices will, however, affect the size of the coefficient for the correlation between the matrices, because of the effect on the standard deviations of the distributions of coefficients for the matrices. Before the computation of the coefficient, we must therefore choose a criterion for the direction of the variables. Two criteria have been applied. First we chose the correlation of the variables with sociometric status as a common criterion. Then the "positive" direction is defined through a positive coefficient for the correlation between the

ratings in the variable and sociometric status. Secondly, some of the rows and the columns in the matrix were reflected, so as to give only positive sums för the columns. (For procedure see e.g. Fruchter, pp. 64–66). Reflection of the variables 1, 7, 13, 14 and 19 fulfilled the claim and gave the required matrix. This procedure gives the same direction for all the variables as that obtained by the first criterion.

The coefficient for the correlation between the two matrices was .942. The coefficient must, however, be cautiously interpreted, since its size depends, among other things, on the relevance of our criteria for the direction of the variables and on the intercorrelations in each distribution of coefficients.

B. Cluster analysis

A cluster analysis was performed as a crude estimation of the structure in the 16×16 correlation matrix for comparison with the structure found by the factor analysis. (For a theoretical discussion of cluster analysis, its limitations and its relation to factor analysis, see e.g. Cattell, 1957, pp. 14–33. For the practical method of estimating clusters see Holzinger 1941, pp. 21–28. The cluster analysis was performed by an assistant who had no other contact with the investigation that could affect her finding of the variables for each cluster. The factor analysis and the cluster analysis were thus performed quite independently of each other.) We have chosen to define a cluster as those variables in the matrix, whose lowest intercorrelation coefficient is significant at the 5 per cent level. This means, in our case, that the lowest intercorrelation coefficient should be .25 in any cluster.

The clusters, found in this way, are reported here, together with the factors to which they best correspond. For the variables are given the mean correlation coefficient with the other variables in the cluster and the factor loadings.

We can immediately by inspection find rather close agreement between the variables in clusters I and IV and factors A_1 and C_1. The variables in factor B_1 are divided into the two clusters II and III, which are partly correlated, and give what Cattell calls a nuclear cluster (Cattell, 1946, p. 236). Generally one also finds more clusters than factors. Of the variables with high loadings in

TABLE 9. *Clusters for the students' ratings. Corresponding factors and factor loadings*

Cluster			Variable			Mr	Factor	Factor loading
I	1	(6)	Dull	—	Acute	.63	A_1	.74
	2	(11)	Dependent	—	Independent	.57		.71
	3	(18)	Lacking vitality	—	Vital	.48		.49
	4	(13)	Subjective	—	Objective	.47		.56
	5	(14)	Anxious	—	Confident	.43		.64
	6	(10)	Insensitive	—	Sensitive	.36		.39
II	1	(1)	Intropunitive	—	Extrapunitive	.55	B_1	.49
	2	(19)	Tolerant	—	Intolerant	.54		.23
	3	(13)	Objective	—	Subjective	.49		.00
	4	(9)	Scrupulous	—	Unscrupulous	.39		.20
	5	(17)	Warmhearted	—	Cold	.39		.50
III	1	(3)	Submissive	—	Dominant	.57		.40
	2	(2)	Reticent	—	Ostentatious	.55		.19
	3	(15)	Unsociable	—	Sociable	.43		.04
	4	(1)	Intropunitive	—	Extrapunitive	.42		.49
	5	(4)	Believing in authority	—	Oppositional	.38		.50
	6	(9)	Scrupulous	—	Unscrupulous	.37		.20
IV	1	(12)	Gloomy	—	Cheerful	.54	C_1	.90
	2	(15)	Unsociable	—	Sociable	.47		.88
	3	(7)	Maladjusted	—	Well-adjusted	.44		.57
	4	(14)	Anxious	—	Confident	.27		.27

factor D_1, only 9, Scrupulous, is included in the 16×16 variables matrix.

The coefficient for the correlation between the matrices is .942. Even if this coefficient should not be taken as an exact value for the agreement, it may—together with the result of the cluster analysis—be considered as strong support to the assumption that the factor analysis of the first matrix can account for the variance in the second matrix too. Therefore we have seen no sufficient reason to perform a new factor analysis for the second matrix, especially as the main purpose of the factor analysis was to investigate whether there is any structure in the matrix from the students' ratings, and see if a possible structure is meaningful

PART II

Projective test data

Chapter 4. The collecting of the protocols

1. SUBJECTS

With one exception all male students at four schools in the investigation served as subjects in the TAT-test sessions (see p. 35). The only exception was one student, who, for reasons of principle, did not take part in any tests. As there were 65 students at the four schools, the TAT was administered in individual test sessions to 64 individuals. The recording of one of the protocols did not succeed for technical reasons. Thus the total number of protocols in the TAT investigation is 63. The number of protocols in various groups is given in Table 10. As was pointed out earlier, the students at one school were divided into two groups. This explains the five groups of individuals.

TABLE 10. *Number of protocols in various groups*

Group	I	II	III	IV	V	Total
Protocols	13	17	10	13	10	63

2. PICTURES

The original series of pictures for the TAT consisted of 20 pictures. The pictures were intended to be administered in two sessions with ten pictures in each session (Murray, 1943, pp. 3–5). In accordance with common practice nowadays, we chose to give the test in one session, with a sufficient number of pictures to give as much information as possible. (See e.g. Vernon, 1957, p. 181, Bellak, 1954, p. 100). Fifteen pictures were chosen and administered in the following order: 1, 2, 3 BM, 4, 5, 6 BM, 7 BM, 8 BM, 10, 11, 12 M, 13 MF, 14, 18 BM and 16. In accordance with the instruction the students were allowed five minutes for each picture. If five minutes had been used for each picture it would have

given a total time of 75 minutes. Actually only very few subjects required more than 60 minutes to tell the stories of the fifteen pictures.

3. TEST SESSIONS

The TAT-pictures were given individually. The same psychologist administered the test to all the subjects.

A. Instruction

Before the first picture was handed over the following instruction was given:

"This is a test of imagination. I am going to show you some pictures, one at a time and your task will be to make up a story for each. The picture you see on the card can be used as an illustration to the story you tell.

"Tell what has led up to the event shown in the picture, describe what is happening at the moment and then give the outcome. Don't forget to tell about the feelings and thoughts of the characters.

"Speak your thoughts freely as they come to your mind. You can tell any story. There is no story that is right or wrong.

"You can devote about five minutes to each story."

After the instruction the tester's comments were limited to:

1. A short repetition of the instruction after the first story if required.

2. A question about the outcome of the story, if the subject did not tell any.

The instruction agrees in the main with that proposed by Murray (1943, pp. 3–4). The most important deviation is the following: Murray proposed that the instruction should begin: "This is a test of phantasy, one form of intelligence." In our instruction we excluded the statement about the phantasy production as an indicator of the subject's intelligence (see Bellak, 1954, p. 38).

The inquiry after the stories was limited to an endeavour to find out how much autobiographic material was included in the stories.

Immediately after finishing the stories, the subject was in-

structed to divide the pictures into two groups according to the degree of spontaneous liking for the pictures. The number of liked and disliked pictures was not fixed. Finally, each subject placed the pictures in order of preference.

B. Recording

All the stories were recorded on a tape recorder, which was openly shown to the subject. The apprehension felt before the investigation, that this way of recording would restrain the subjects in their test behavior was not verified on the observable level of behavior. Several subjects said that they had completely forgotten the tape recorder during the test session.

For each story the initial time was recorded. This time was taken from the moment the picture was handed over, to the moment the subject began to speak. The total time for each story was also recorded.

All the stories were type-written from the tapes. Every pause in the story, lasting for more than five seconds, was recorded with a stop-watch. These pauses were indicated in their appropriate places in the protocols.

The final protocols, in the form in which they were judged by the test examiners, thus contained the stories told by the subjects, all the comments made by the tester during the stories, initial time, total time and pause times as reported above.

C. Test length

The mean number of words per story was 148, and the standard deviation for the distribution 86.2.

A direct relation between test length and possibility of interpretation was stated by Murray in his manual (1943, p. 4). He stated that three hundred words is the standard average length of stories by adult subjects and he gives this as a norm for the ability of the examiner to perform his function. The claim is probably seldom fulfilled. Accounts of investigations with the TAT show that the mean number of words per story usually lies between one and two hundred. An investigation published in 1954 may serve as an example (Milam, 1954). Three groups of uni-

versity students, who thus form a positively selected material, were given the TAT, during which the examiner assumed different attitudes. The means in the groups were respectively 113, 160 and 163 words per story.

Bellak has pointed out that brevity need not be a serious drawback (Bellak, 1954, p. 40). From an investigation with the TAT Hartman reports: "A review of the TAT records also revealed that the subjects most successfully judged were the least fluent in their TAT-responses, which supports W. E. Henry's contention with reference to the validity of interpreting short TAT records." (Hartman, 1949, p. 30, Henry, 1947, 35, pp. 113–5.)

The number of words from individuals in one cultural setting need not necessarily be the same as that from individuals in another, since cultural factors probably affect the size of both the passive and active vocabularies, as well as other language customs (see Carlsson, 1955, 269–78). Our results do not, however, indicate any serious differences from the usual number of words per story. Other Swedish investigations for comparison of test length are not available.

Chapter 5. Judging of the protocols

1. TEST EXAMINERS

To judge the test protocols four psychologists were engaged. In the following they are called A, B, C and D.[1]) Each of the examiners has rated each protocol in all the variables.

When engaging test examiners in the investigation we found that only a limited number of clinical psychologists in Sweden used the TAT in such a systematic way that they could act as test examiners in the investigation. Therefore, those who undertook the task had different training and experience of the test. Before the beginning of the investigation examiner A had taken and judged 75–100 TAT protocols, examiner B 150–200, examiner C more than 500 (always together with other projective tests) and examiner D, whose speciality is Rorschach, some tens of TAT protocols. All the examiners have had experience of similar techniques, and have used them in clinical work.

The four examiners all have academic degrees with psychology as one subject. They all take part in clinical work, where they use the TAT. None of the examiners participated in any other way in the investigation and had no contact with the material other than in judging the protocols. The examiners have also worked quite independently of each other.

2. VARIABLES

The construction of the variables was performed in collaboration with psychologists with experience of the TAT in clinical use. This is accounted for earlier (p. 25). The definition of the final set of variables common for the four tests in the investiga-

[1]) From the beginning two further test examiners were engaged. Since they have rated only two groups of students each, their ratings have not been included in the calculations reported here.

tion was however partly the result of a compromise between the definitions of the variables for different tests. For several variables the definitions were not identical, but so alike that there were not sufficient reasons to use more than one. The one used was then formulated so as to cover what the various tests were expected to give information about. Therefore, before the ratings began, the whole final set of variables was discussed with TAT experts, who judged which of the variables were possible to use in the TAT investigation. From this discussion the result was that four variables were excluded from the total set of twenty variables for the ratings based on the TAT. The variables excluded were: 5, Impulsive, 8, Careful, 16, Reflective and 20, Emotional.

In the TAT investigation were then included the following sixteen variables:

1. Extrapunitive
2. Ostentatious
3. Dominant
4. Oppositional
6. Acute
7. Maladjusted
9. Scrupulous
10. Sensitive
11. Independent
12. Cheerful
13. Subjective
14. Anxious
15. Sociable
17. Warmhearted
18. Vital
19. Intolerant

The above definitions are those with scale value 7.

3. THE TEST EXAMINERS' RATINGS OF THE PROTOCOLS

The examiners' ratings of the subjects were performed on the same type of form as was used by the students during their ratings of each other. The only difference was that the names of the students were replaced by the corresponding code numbers.

A. Instruction

a. *Prediction of the students' ratings.* The instruction to the test examiners was to rate the protocols in the group in a normal distribution in one variable at a time with the same technique that

the students used in their ratings of each other (see p. 27). The ratings were made for one group at a time on all variables before the next group was rated. The ratings should be performed so as to agree with the ratings made by the students. The task was thus to predict the students' ratings of each fellow in each variable on the basis of the test protocol.

Two important aspects of the instruction may now be considered. First, the task for the examiners involves at least two more steps than the usual clinical procedure in judging a personality from a TAT protocol. The task is to judge how the personality, reflected in the protocol, manifests itself in observable behavior and how this behavior is observed and judged by the other individuals of the group. These two last steps are not included in the usual judging procedure. Second, the criterion ratings are made by untrained observers. We may assume that those ratings are expressions of overt behavior without much interpretation of the latent psychological variables that may generate the observable behavior.

b. *Rating of latent behavior.* As mentioned earlier there has been much discussion about the relation between manifest and latent behavior. The examiners were therefore given an opportunity to express a possible difference in the positions of the individuals on the continua for manifest and latent behavior. In those cases where the test examiners judged the positions of an individual to be different, he could mark this by two ratings, one for each position.

c. *Rating of subjective confidence.* Some of the variables might be more difficult than others to judge from the protocols. Greater difficulties may be encountered in some protocols than in others. In order to study the examiners' subjective experiences of the variations, they were instructed to mark the degree of subjective confidence with which each prediction of the students' ratings was performed. These ratings of subjective confidence were made on a four point scale from 0 to 3, where 3 meant maximum subjective confidence. No instruction was given for the distribution of the ratings of subjective confidence.

d. *Rating of one variable at a time.* The instruction to the students emphasized the importance of the ratings being made in one variable at a time over all individuals in the group (p. 27).

This was done in order to minimize a possible influence of the halo-effect. The same instruction was given to the test examiners.

B. Material

In addition to the protocols in the form described earlier (p. 55), the test examiners were given the following information about the individuals:

1. age and sex, 2. ages of the siblings, 3. liked and disliked pictures, and 4. the rank order for liked and disliked pictures.

C. Scoring system

As was stated in the introduction, the investigation is explorative in character. No hypotheses were stated before the investigation. For these reasons no special scoring system was fixed for the test examiners. They were free to use the method they had found most useful.

When all the ratings had been performed, the examiners reported which scoring method they had followed. Examiner A followed a modified form of Murray's need-press analysis with interpretation of the constellation of hero, environment, theme, feeling and outcome, special consideration being taken to the conditions given by the variables defined (Murray, 1938). The method used by examiner B agrees most with that proposed by Tomkins with special stress on the analysis of themes (Tomkins, 1947). Examiner C usually follows the method proposed by Aron (1949) and it has been used for the ratings by C in the present investigation. Examiner D, who earlier used Murray's need-press analysis in his clinical work, has followed the more intuitive technique suggested by Stein in the revised edition of his manual (Stein, 1955).

Chapter 6. The test examiners' ratings

1. GENERAL REMARKS

The test examiners reported considerable difficulties in rating one variable at a time for all individuals in a group before proceeding to the next variable. Three of the examiners (A, B, D) reported that they found the instruction so difficult in this respect, that they rated all the variables for one subject before going to the next, consideration being taken to the fact that the individuals should be compared with the group standard, and that the distribution of ratings should be normal. Examiner C rated three groups according to the instruction and the remaining two groups in the same way as the other examiners.

2. TRANSFORMATION OF THE EXAMINERS' RATINGS

In order to make possible the use of all the students' ratings for each variable in one calculation, the distributions of ratings from various groups were transformed to the same mean (15.5) and the same standard deviation (5.0) (p. 30). The same transformation was performed for the test examiners' ratings. The coefficients of correlation based on these transformed values and calculated over all individuals at a time will thus give the correlation within groups with N—m—1 degrees of freedom, where m is the number of groups (Lindquist, 1940, pp. 221–2).

3. AN ANALYSIS OF THE TEST EXAMINERS' RATINGS

A. Problems

In connection with the account of the students' ratings, the results of a factor analysis (for the 20×20 matrix) and a cluster

analysis (for the 16×16 matrix) were given. In this chapter an analysis of the structure in the matrices for the test examiners' ratings is performed. The chapter ends with an attempt to compare the matrices of subjects and test examiners, according to structure.

The study of the structure in the ratings by the students and the test examiners is an important problem from both theoretical and practical points of view. It is reasonable to consider the structure of a matrix to some extent as the result of and an expression of a more or less conscious theory on the relation between various variables in the personality system. If the examiners have the same theory in this meaning, it will result in an agreement in the structure of the ratings, quite independent of the degree of inter-rater reliability. It is possible to obtain identical correlation matrices for two examiners, even if the coefficients for the correlation between the two examiners' ratings is zero in each variable in the matrix. (It is, however, not possible to obtain different matrices, if the inter-rater reliability is 1.0 in each variable. Then, of course, the correlation matrices must be indentical.)

Another application of the comparison between the structure for the test examiners mutually and between test examiners and students, is the relation between the degree of agreement in structure and the size of the validity coefficients for various examiners. Does there exist a positive relationship, so that an examiner with great agreement in structure with the students' ratings will obtain higher validity with the students' ratings as a criterion, than another examiner with less agreement?

B. Calculations

The coefficients of the correlation matrix for each of the examiners were computed over all subjects at a time on the transformed scores. The correlation matrices are given in Tables 52, 53, 54 and 55 in Appendix A (pp. 156–159).

In order to facilitate further computations and comparisons we defined a common direction for the variables. The direction chosen will determine the signs of the coefficients in the matrices. We chose to determine the direction of the variables with the same two criteria as those discussed and used in an earlier computation

(pp. 48–49). The direction, determined by reflection of some variables of the matrix for the students' ratings and by the positive correlation with sociometric status for these ratings, has been used throughout for the variables of the test examiners' matrices, except when otherwise stated.

C. Cluster analysis

a. *General discussion.* For the study of the structure in the matrices for the examiners' ratings, it was possible to use two principal methods, factor analysis and cluster analysis. Both the methods have their limitations. The factor analysis is time-consuming, and the subjective character of the rotation of the axes will yield factor matrices, whose differences may be attributable to either actual differences in the factor structure, or to the subjective moment of rotation, even if there does not exist any true differences in the structure. This last effect is to be expected, especially when the data, upon which the correlation matrices are based, are ratings of the type used in this investigation and when the number of individuals is relatively small. The cluster analysis is easier to perform and takes considerably less time. However, the cluster analysis is—both in determining the criteria for the clusters, which will affect the boundaries of the clusters, and in determining which of two variables, not intercorrelated, shall next be included in the cluster—even more subjective in character than the factor analysis. Nor does it give anything but the superficial grouping of the variables, which correspond to the syndromes of the matrix, not to the real generating factors. (For a general discussion on cluster analysis and factor analysis see e.g. Cattell, 1957, pp. 14–33.)

In our choice of method for an analysis of the structure of the examiners' ratings, we had to weigh the two alternatives: factor analysis and cluster analysis. The advantages of the former—less subjectivity, obtaining factors instead of syndromes—were found, however, to be outweighed by the disadvantages of the considerably longer time required. Another reason for the choice was that no exact and fairly simple method of determining agreement between factor matrices is as yet available (Henrysson, 1957, p. 121). We chose, therefore, the method of cluster analysis.

The agreement between the matrices is thus limited by the method used to the relatively superficial aspects that can be illuminated by cluster analysis. Together with the other results in the comparison, it may however throw some light on the structure of the matrices and be used for some tentative interpretations of the agreement found.

b. *Definition of cluster.* The cluster analysis has been performed as for the cluster analysis of the students' ratings. As in that analysis we have defined a cluster as those variables whose lowest intercorrelation coefficient is significant on the 5 per cent level, which means a coefficient of .26. (In some clusters has been included one variable with one intercorrelation coefficient of insignificant but near significant size. This has been done so as to give the best possible solution. This is the case in one cluster for examiners A and B and in three clusters for C and D.)

c. *Clusters.* The clusters found are given in Appendix A (Tables 56, 57, 58, and 59, pp. 160–163). The mean coefficient for the correlation with other variables in the cluster is given for each variable.

For examiner A four clusters were obtained. Five variables were included in more than one cluster. For examiner B the analysis gave three clusters. Four variables (7, Well-adjusted, 12, Cheerful, 13, Objective and 14, Confident) are included in two clusters, forming a nuclear cluster. The matrix for examiner C gives two highly correlated clusters and one smaller cluster, consisting of four variables. In the first two clusters, seven variables form a nuclear cluster. The same four variables as for examiner B are included there. For examiner D five clusters were obtained. Two variables, the same as those included in the nuclear clusters for examiners B and C (12, Cheerful and 14, Confident), contribute to two clusters.

d. *Comparison of clusters for the test examiners.* There is no exact method for comparing the results of cluster analysis. Such an exact method would also be more or less meaningless on account of the method of obtaining the clusters. Therefore it is impossible to draw any exact conclusions from the results of the cluster analysis in our case. We have, however, tried to make a comparison.

There is no general agreement between the matrices for the in-

dividual examiners. Some tendencies to agreement are, however, rather obvious. In order to illustrate these tendencies we have assembled some clusters in three groupings.

The first grouping is given in Table 11.

TABLE 11. *Grouping of clusters for the examiners' ratings. I.*

Variable	A_{II}	B_I	C_{II}	D_I
1.				
2. Ostentatious	.54	.48	.45	.41
3. Dominant	.51		.47	
4.				
6. Acute		.53	.59	
7. Well-adjusted		.48	.57	
9.				
10.				
11. Independent	.56	.50	.42	
12. Cheerful	.45	.45	.51	.41
13. Objective		.46	.56	
14. Confident	.45	.57	.56	.42
15. Sociable	.53	.55	.64	.43
17.				.46
18. Vital	.62	.50	.58	.45
19.				

For five variables all the examiners agree about their inclusion in the cluster. For one variable three examiners and for four variables at least two raters agree about the inclusion of each of the variables in the same cluster as the five main variables.

The next grouping of clusters is given in Table 12.

Here we find a rather good agreement in the content of the clusters. The four raters agree in including variables 7, 12, 13 and 14. Three of them also agree about variable 19. For a tentative interpretation the variables are best given by their "negative" direction: Maladjusted, Gloomy, Subjective, Anxious and Intolerant. Variable 7, Maladjusted, is the one with the highest mean intercorrelation for all the clusters. This leads to a possible interpretation of the variables forming these clusters as indicating personality conflicts and disturbances. It must be pointed out that variables 7, 12, 13 and 14 form a nuclear cluster for examiners B and C (p. 64). These variables are also included in the clusters in

TABLE 12. *Grouping of clusters for the examiners' ratings. II.*

	Variable	A_{IV}	B_{II}	C_{I}	D_{III}
1.					
2.					
3.					
4.				.45	
6.				.48	
7.	Maladjusted	.52	.67	.65	.51
9.			.55		
10.				.54	
11.					
12.	Gloomy	.47	.58	.56	.49
13.	Subjective	.52	.61	.60	.29
14.	Anxious	.48	.50	.50	.50
15.				.59	
17.				.56	
18.				.61	
19.	Intolerant	.40	.51	.49	

grouping I. For examiners A and D the variables 12 and 14 contribute to two clusters, those in groupings I and II.

The third grouping of clusters is given in Table 13.

TABLE 13. *Grouping of clusters for the examiners' ratings. III.*

	Variable	A_{III}	B_{III}	C_{III}	D_{IV}
1.	Extrapunitive	.56	.46	.52	.37
2.					
3.			.38		
4.	Oppositional	.44	.44		.32
6.					
7.					
9.				.30	
10.	Insensitive	.48		.51	
11.					
12.					
13.					
14.					
15.					
17.	Cold	.52	.40	.53	
18.					
19.	Intolerant	.52			.28

In this constellation all the raters agree only on one variable, viz. 1, Extrapunitive. For variable 4, Oppositional and 17, Cold, at least three raters agree in including the variables in the same cluster as variable 1. For variables 10, Insensitive and 19, Intolerant, two examiners agree.

The clusters may be interpreted in the same way as factor B_1 in the factor analysis, as characterised by the tendency "against" others (p. 46).

All the comparisons and the interpretations of the groupings of clusters must be interpreted in the light of the limitations in the method of cluster analysis, which has been discussed earlier.

D. Agreement between the test examiners on the intercorrelations for one variable at a time

The earlier comparisons between the matrices for the examiners' ratings indicate a tendency towards agreement. Therefore we have made a further computation in order to give a more detailed picture of the relationships. The following procedure has been followed.

a. *Procedure and results.* For each examiner the intercorrelation coefficients with all other variables were rank ordered for one variable at a time. The rank orders were based on the coefficients in the correlation matrix in which the direction of the variables had been determined by the application of the two criteria of reflection and positive correlation with sociometric status.

Thus we obtained sixteen rank orders for each matrix, one for each variable in the matrix, with fifteen ranks in each rank order. For each variable we obtained four rank orders, one for each of the examiners. The degree of agreement between the examiners in the rank orders for each variable was estimated according to Kendall's coefficient of concordance (Edwards, 1954, pp. 402–7). For determining their significance levels, the coefficients were transformed to χ^2-values (Edwards, p. 411). The coefficients, the χ^2-values and the p-values are given in Table 14.

The coefficient of concordance indicates the degree of agreement for the whole set of four rank orders. Another aspect of the problem is how well any one examiner agrees with another exam-

iner. For this we need the mean intercorrelation coefficient between the rank orders for each variable. It is possible to compute this mean rank coefficient directly from the coefficient of concordance (Edwards, 1954, p. 412). The mean intercorrelation coefficients, estimated in this way for each variable are given in Table 14.

TABLE 14. *Agreement between the test examiners as to the rank orders of intercorrelation coefficients in the matrices*

Variable	W	χ^2	p	M_{rho}	p
1. Intropunitive	.808	45.250	< .001	.74	< .01
2. Ostentatious	.824	46.169	< .001	.77	< .001
3. Dominant	.831	46.557	< .001	.77	< .001
4. Oppositional	.760	42.550	< .001	.68	< .01
6. Acute	.553	30.975	< .01	.40	
7. Well-adjusted	.788	44.100	< .001	.72	< .01
9. Scrupulous	.450	25.225	< .05	.27	
10. Sensitive	.460	25.750	< .05	.28	
11. Independent	.707	39.575	< .001	.61	< .05
12. Cheerful	.775	43.400	< .001	.70	< .01
13. Objective	.767	42.925	< .001	.69	< .01
14. Confident	.666	37.275	< .001	.55	< .05
15. Sociable	.638	35.750	< .01	.52	< .05
17. Warmhearted	.521	29.150	< .01	.36	
18. Vital	.580	32.482	< .01	.44	
19. Tolerant	.715	40.050	< .001	.62	< .05

χ^2 sign. at 1 ‰ level = 36.123 Df = 14 Df = 13
1 % level = 29.141
5 % level = 23.685

As in other estimates of agreement by means of the coefficient of concordance in this investigation, we must evaluate the size of the coefficients in the light of the fact that the coefficients on which the rank orders are based are not independent, as is required by the exact application of the coefficient. Even if we use the values with great caution the results point to some conclusions. There are big differences in the degree of agreement for different variables. For some variables the agreement is so great that the probability of our making a Type I error by the rejection

of the null hypothesis is very small. For some variables the χ^2-value is so near significance at the 5 per cent level that it is doubtful if the significance may be attributable to the fact that the ranks have not been independent.

The mean intercorrelation coefficients must also be interpreted with caution for the same reason as the coefficients of concordance. They are nothing but the same fact, given in another form. The coefficients only state the degree of agreement between any two examiners in their rank orders. Here we see more clearly that the agreement for some variables is small.

In order to study the relative agreement between the examiners, the coefficients for the correlation between the rank orders in each variable were computed for every pair of examiners. The mean coefficients are given in Table 15.

TABLE 15. *Mean rank order coefficients for the agreement between the test examiners' matrices*

	A	B	C	D	M_{rho}
A		.62	.48	.63	.58
B			.68	.55	.62
C				.54	.57
D					.57

4. AGREEMENT IN STRUCTURE BETWEEN THE STUDENTS' AND THE TEST EXAMINERS' MATRICES

A. General remarks

The extent to which the structure of the matrices for the students and the examiners agrees has some practical applications. In this investigation it is possible to relate the degree of agreement in structure to the degree of agreement in the individual ratings, as illustrated in the validity coefficients.

Here it may again be pointed out that the comparisons are restricted by the limitations of the methods used for studying the structure. Another factor which may affect the comparison is that the coefficients in the matrix for the students' ratings are based on ratings by more than one student. The structure in this matrix will

thus be the mean structure for the students. The other matrices are only based on ratings by one examiner.

In spite of these limitations, the comparisons have been performed, as they may possibly throw some light on the problems drawn up.

The degree of agreement between the matrix for the ratings made by each of the examiners and the matrix for the students' ratings has been studied in two ways, as has the agreement between the test examiners mutually.

B. Comparison of clusters for examiners and students

Cluster analyses of the matrices for the ratings by the subjects and by each of the examiners have been reported above. The clusters found are given in Appendix A (pp. 160–163).

By inspection of the tables we find that there is no general tendency to agreement between any of the cluster matrices for the examiners and the cluster matrix for the students. We can, however, find a tendency to agreement between examiners and students in single clusters. We have earlier accounted for three groupings of the clusters for the examiners. One of these group-

TABLE 16. *Grouping of clusters for examiners and students*

Variable	A_{IV}	B_{II}	C_{I}	D_{III}	$Students_{IV}$
1.					
2.					
3.					
4.			.45		
6.			.48		
7. Maladjusted	.52	.67	.65	.51	.44
9.		.55			
10.			.54		
11.					
12. Gloomy	.47	.58	.56	.59	.54
13 Subjective	.52	.61	.60	.29	
14. Anxious	.48	.50	.50	.50	.27
15. Unsociable			.59		.47
17.			.56		
18.			.61		
19. Intolerant	.40	.51	.49		

ings gave close agreement for clusters containing variables which indicate personality conflicts and disturbances (see pp. 65–66.) In Table 16 are given the clusters in this grouping for the test examiners and the corresponding cluster for the students.

As seen there is a clear tendency to correspondence between the test examiners and the students about this cluster. For three variables both examiners and students agree in including them in the cluster. This finding will be used in a later connection.

C. Agreement between examiners and students on the intercorrelations for one variable at a time

For a further and more careful study of the agreement between the matrices for the examiners' ratings and the matrix for the students' ratings we followed the same procedure as in the comparison between the examiners' matrices (see p. 67).

TABLE 17. *Agreement between each examiner and the students on the intercorrelations for one variable at a time*

Variable	A	B	C	D	M_{rho}
1.	.76	.71	.66	.42	.64
2.	.81	.59	.63	.54	.64
3.	.87	.74	.59	.76	.74
4.	.79	.60	.66	.78	.71
6.	.29	.11	.42	.44	.32
7.	.66	.57	.74	.84	.70
9.	.20	−.20	.48	.39	.22
10.	.22	−.18	−.12	.65	.14
11.	.32	.43	.53	.60	.47
12.	.62	.36	.51	.69	.55
13.	.16	.05	.11	.03	.09
14.	.55	.64	.72	.30	.55
15.	.60	.49	.35	.49	.48
17.	.58	.51	.40	.10	.40
18.	−.37	−.15	−.01	.02	−.13
19.	.74	.24	.51	.53	.51
M	.49	.34	.45	.47	

Sign. level for 5 % = .52 Df = 13
1 % = .65
1 ‰ = .77

For each variable in the matrix of the students' ratings the intercorrelation coefficients were rank ordered. The degree of agreement between these rank orders and the corresponding rank orders based on the intercorrelation coefficients of the test examiners' matrices was computed for each examiner and each variable by means of rank order coefficients. In this way we obtained sixteen coefficients for each examiner, which are given in Table 17. The significance levels have been determined according to McNemar (1955, p. 210).

From the table we can draw the conclusion that there is no general agreement between the matrix for the students and the matrices for the test examiners, when the comparisons of the matrices are performed for one variable at a time in this way. For some variables the degree of agreement is very small. In one case the mean coefficient is negative.

The degree of agreement in the whole set of coefficients in Table 17 was estimated in a coefficient of concordance for the rank ordered coefficients (Edwards, 1954, pp. 402–7). The coefficient = .707. This gives a χ^2-value of 42.397, which is significant at the 1 per cent level (Df = 15). Even if this value must be interpreted with caution, it is indicative of substantial agreement. The reliability of the average rank order is good (.87).

The mean rank coefficients for the agreement on the intercorrelations for one variable at a time between each of the test examiners and the students are reported in Table 18. The coefficients are given in order of size. For each variable there is also reported the mean rank coefficient for the agreement in structure between any two test examiners (see Table 14, p. 68).

We find a marked tendency towards agreement between the rank orders of the coefficients in the table. The only variables for which the coefficients are very different are variables 13, Objective, and 18, Vital. This result indicates that the examiners are in mutual agreement as regards structure for the same traits in which agreement exists between them and the students. And they disagree with each other regarding the same traits about which they disagree with the students.

The mean level of the agreement with the students' matrix for various test examiners is indicated by the coefficients of the last

TABLE 18

Variable	M_{rho} (students/examiner)	M_{rho} (examiner/examiner)
1. (3) Dominant	.74	.77
2. (4) Oppositional	.71	.68
3. (7) Well-adjusted	.70	.72
4. (2) Ostentatious	.64	.77
5. (1) Intropunitive	.64	.74
6. (14) Confident	.55	.55
7. (12) Cheerful	.55	.70
8. (19) Tolerant	.51	.62
9. (15) Sociable	.48	.52
10. (11) Independent	.47	.61
11. (17) Warmhearted	.40	.36
12. (6) Acute	.32	.40
13. (9) Scrupulous	.22	.27
14. (10) Sensitive	.14	.28
15. (13) Objective	.09	.69
16. (18) Vital	–.13	.44

line of Table 17. For comparison, they are given together with the mean coefficients for the agreement between all the test examiners' matrices, given earlier (Table 15, p. 69).

	A	B	C	D
Agreement with other test examiners	.58	.62	.57	.57
Agreement with the students' ratings	.49	.34	.45	.47

It will be noticed that the mean level is higher for the agreement in structure between the test examiners' matrices than for the agreement between the matrices for the test examiners' ratings and the matrix for the students' ratings. The same tendency is reflected in the coefficients in Table 18.

One other tendency is worth mentioning. One expert (B) differs from the other experts in his mean coefficients. For the mutual agreement as regards structure between the test examiners his coefficient is the highest. For the agreement with the matrix for the students' ratings he has the lowest coefficient. The difference in the last case is marked.

5. SOURCES OF VARIANCE FOR THE RATINGS BY THE STUDENTS AND THE TEST EXAMINERS

Before leaving the account of the rating data it seems relevant to discuss a further aspect, namely the possible sources of the common factor variance for the ratings by the students and the test examiners.

A. The variance of the students' ratings

The common factor variance in the correlation matrix for the students' ratings may be considered as mainly derived from the following sources:

a) the objective degree of occurrence of the trait in observable behavior,

b) what Guilford calls "the logical error", "the apparent logical coherence of various traits irrespective of individuals" (Guilford, 1954, p. 279), and

c) the halo effect.

In the students' ratings the logical error may be expected to be very small, since the ratings were performed for one variable at a time and with the direction of certain variables reversed. Our earlier calculations indicate that the halo effect may play some role but is minimized by the precautions taken. Most of the common factor variance may be expected to be due to the objective degree of occurrence of the traits in observable behavior.

B. The variance of the test examiners' ratings

The common factor variance in the correlation matrix for the test examiners' ratings may be considered as mainly due to the following sources:

a) the objective degree of occurrence of the trait in the individuals rated,

b) the logical error, which may reasonably be the result of a more or less conscious personality theory for the test examiners and

c) the halo effect.

In the test examiners' ratings a) is the main part of the variance which gives the validity. Halo may play its role in some form, based on any irrelevant source, such as the test length (see later ch. 13). Much of the variance may reasonably be due to the "logical error", as it is defined above. It seems reasonable to assume that the test examiners, when uncertain about a rating, incline towards conclusions from earlier ratings of the same individual in other traits. The "logical error" may be expected to play a relatively important role when the ratings are performed in blind analysis and with the ratings delivered for one individual at a time for all variables.

6. RATINGS OF LATENT BEHAVIOR

A. General remarks

The task for the examiners, given in the instruction, was to predict the ratings given by fellow students for each individual. The predicted ratings were based on observations of manifest behavior. Thus the task was to predict how the latent tendencies of the individual expressed themselves in behavior and then to predict how this behavior was observed and rated by the fellow students. In discussions before the beginning of the investigation, it was pointed out that for some individuals there might be a difference between the individual's tendency towards a certain type of behavior and the actual expression of this same tendency. Therefore the test examiners were given an opportunity to express a judged difference between manifest and latent behavior by making two ratings, one for each (see p. 59).

Only one examiner (C) used this possibility consistently. This will diminish the use of this type of ratings. Here will be given, however, some of the relations of these ratings by C to the predictions by C of the students' ratings.

B. The relation between ratings of latent and manifest behavior from the TAT

The coefficients for the relationship between the ratings of

latent behavior by C and his ratings of manifest behavior are given in Table 19.

TABLE 19. *Correlation between ratings of latent and manifest behavior (Examiner C)*

Var.	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
r	.32	.49	.69	.56	.93	.82	.84	.86	.82	.92	.87	.74	.85	.82	.96	.98

N = 63 Df = 57

The median of the coefficients is .83.

Already the obtained correlation coefficients give a clear tendency to agreement in most variables between ratings of manifest and ratings of latent behavior. Only for the variables 1, Extrapunitive, 2, Ostentatious and 4, Oppositional, may the coefficients be considered as low. We must, however, remember that the two types of ratings have been performed at the same time for each subject. This may have had the effect of raising the size of the coefficients.

Another way of presenting the same data, which are the basis for the correlation coefficients may be illustrative. In Table 20 is given the number of students for each variable for which two ratings have been given, indicating a difference in the individual's positions.

TABLE 20. *Number of individuals with different ratings for manifest and latent behavior*

Variable	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
Number	21	18	13	22	3	13	10	9	11	6	3	8	10	7	2	2

Another illustration is Table 21, where are given the frequencies for the different numbers of times that two ratings have been delivered for a student. The first column thus means that there are only nine individuals for whom no double ratings are given.

TABLE 21. *Frequencies of double ratings*

Number of double ratings	0	1	2	3	4	5	6	7	8	9	
Frequencies	9	10	19	8	10	3	1	3	0	1	= 63

7. RATINGS OF SUBJECTIVE CONFIDENCE

For each rating of an individual's position in a variable the test examiner also delivered a rating of the subjective confidence with which the prediction was made. The ratings of subjective confidence were given on a scale 0–3.

The means and standard deviations of the ratings of subjective confidence for each test examiner and variable were calculated. The means of all the ratings of subjective confidence for each examiner were 1.782 (examiner A), 1.837 (examiner B), 1.696 (examiner C) and 1.758 (examiner D).

PART III

Results

Chapter 7. Inter-rater reliability

1. GENERAL DISCUSSION

All the conventional methods of assessing test reliability except the construction of parallel sets of pictures have been applied in attempts to estimate the reliability of ratings based on the TAT. The test-retest method has been used by e.g. Tomkins (1947, p. 6) and Lindzey and Herman (1955, pp. 36–42), re-rating of the same protocols after a time interval by Sen (1953, pp. 91–100) and others, and the split-half technique by estimating the degree of agreement between ratings based on half of the pictures in a TAT series and ratings based on the other half by Lindzey and Herman (1955) among others. The relevance of various methods of estemating the reliability has been discussed by e.g. Tomkins (1947, pp. 4–9), Rotter (1949, pp. 205–6), Lindzey and Herman (1955) and Ås (1958, p. 53).

The estimating of the reliability in ratings based on the TAT has some specific problems. In conventional, standardized ability tests it is possible to hold such factors as the role of the tester, the administration of the test, registering, scoring and the comparing with norms under control to a great extent, so that possible errors from these sources can be minimized. This is done by careful instructions and the construction of strict norms. Thus the role of the tester is minimized, and it is possible to get reliability estimates in which the main sources of error lie in the test itself. The scoring procedure in particular may be kept under control so that the error from this source in most ability tests may be considered as zero or near zero. This is possible in this type of test, because the scoring usually consists of evaluating one answer to one question at a time as either right—in some cases as more or less right—or wrong. For this procedure it is possible to give rules that cover every possible answer. This leads to the same score, independent

of which rater is judging the test protocol. The test scoring is more or less automatic.

In most projective techniques, e.g. the TAT, the situation is different. For the "scoring" of a projective test we cannot give rules that cover every possibility of interpretation of the stimulus material by the tester. Single signs are seldom if ever used as a basis for a rating; usually we use a constellation of signs. Neither do we sum up single values to approximate a "true" score for the individual's position on the variable. These circumstances mean that a great subjective component is included in judgments, based on the TAT, as well as on most projective methods. The variance in a distribution of scores from a certain examiner for a group of individuals is to a high degree due to the examiner's conception of the interpretation of the test material. How much this source of subjectivity in the scoring procedure will affect the reliability depends among other things on the scoring method used.

Probably influenced by the great role of the subjective component in the reliability of most rating procedures, most investigators have pointed to the inter-rater reliability coefficient as the best estimate of the reliability (see e.g. Guilford, 1954, p. 395 and Ås, 1958, p. 53). By this we obtain a measure of the degree to which the source of subjectivity in the scoring contributes to the error variance in the distribution of test ratings. It is often neglected, however, that this will be the only error component included in the error variance in inter-rater reliability coefficents, computed for data as in this investigation, when the data are ratings based on only one sample of behavior. In this case other error components, included in the calculation of usual test reliability coefficients, do not contribute to the test variance.

This problem has been taken up and discussed by Cattell (1957, pp. 342–52), who introduces a new term, "conspect reliability", estimated in a conspect coefficient, for the inter-rater reliability. After having mentioned the various attempts to get a standard nomenclature of reliability, he discusses his own propositions about the use of the terms consistency, stability, equivalence, and dependability coefficients as reliability coefficients for various methods of estimating reliability. He summarizes his discussion about

the possible sources of error variance in various coefficients by splitting the error variance into the following components:

σ_{eo}^2 = conspect error: the difference of scoring from the view points of two different observers
σ_{ea}^2 = administration error: the difference of two experimenters in the presentation of the test
σ_{ex}^2 = experimental error: variations of response due to chance variations in conditions with the same experimenter, and to ambiguities in the test
σ_{ff}^2 = function fluctuation in the main required factor
σ_{fs}^2 = function fluctuation in the specific factor

The translation of these error sources into correlation coefficients is easily achieved. For the inter-observer reliability coefficient, the conspect coefficient (r_{rc}), we obtain:

$$r_{rc} = \frac{\sigma_f^2 + \sigma_s^2 + \sigma_{fs}^2 + \sigma_{ff}^2 + \sigma_{ex}^2 + \sigma_{ea}^2}{\sigma_f^2 + \sigma_s^2 + \sigma_{fs}^2 + \sigma_{ff}^2 + \sigma_{ex}^2 + \sigma_{ea}^2 + \sigma_{eo}^2} = \frac{\sigma_t^2 - \sigma_{eo}^2}{\sigma_t^2}$$

(σ_f^2 = the required source trait variance. σ_s^2 = the specific variance. σ_t^2 = the total variance.)

As Cattell points out, we can now estimate the fraction of the total variance which is due to observer scoring differences as

$$\frac{\sigma_{eo}^2}{\sigma_t^2} = 1 - r_{rc}$$

It will be clear from the discussion above that the only source of error, which will contribute to the error variance in inter-rater reliability coefficients for ratings based on the TAT-protocols, collected in one test session, is what Cattell calls the conspect error, the difference of scoring from the view points of two different observers.

Attention has been drawn earlier to the fact that the error variance obtained through differences in scoring from two independent observers is very small in usual ability tests, often so small that it may be neglected. In the inter-observer reliability coefficients for ratings based on the TAT is not included such error

variance as that from administration or experimental error (see e.g. Milam, 1954, pp. 221–6, Eron and Ritter, 1951, pp. 55–61, Garfield, Blek and Melker, 1952, p. 140, Cronbach, 1956, p. 175) or from function fluctuations (see e.g. Tomkins, 1947, pp. 4–6, Crandall, 1951, pp. 400–3). These sources contribute to nearly all the error variance which determines the deviation from 1.00 for the reliability coefficients in usual ability tests. In ratings, e.g. ratings from the TAT, these sources of error are less under control and will thus contribute even more to the error variance in the distribution of test ratings. From this point of view the conspect coefficient, which is the estimate of inter-rater reliability, in this case may be considered as a *maximum* value for the reliability of the ratings which are based on the TAT. If the other sources of error were also included in the calculation of the reliability coefficients, we should obtain still lower coefficients than for inter-rater reliability only.

2. THE PRESENT INVESTIGATION

A. Statistical procedure

The reliability of the ratings made by the experts has been computed by the intraclass formula proposed by Ebel (1951, pp. 407–24). It is possible with this method to get a convenient value of the exactness of the reliability coefficients.

The reliability coefficients in our case are computed over all individuals on the transformed scores (p. 61). This means that the means and the standard deviations are the same for the four examiners. Thus in the use of the intraclass formula for estimating the reliability, the variance between examiners is excluded. If the ratings had been used to decide the relative position for every individual in relation to a fixed value it would have been appropriate to include this variance. The method used will give the reliability, when we are interested only in the agreement between the rank orders of the subjects, given by various examiners.

B. Results

In Table 22 the intraclass coefficients for the reliability are reported.

TABLE 22. *Intraclass coefficients etc. for inter-rater reliability*

Variable	$\bar{r}$	$r_{l\ (5\ \%)}$	$r_{u\ (5\ \%)}$	r_{44}	$n_{.80}$	$n_{.90}$
1. Extrapunitive	.22	.10	.33	.53	15	32
2. Ostentatious	.33	.20	.43	.66	8	18
3. Dominant	.17	.06	.28	.45	20	44
4. Oppositional	.34	.21	.44	.67	8	17
6. Acute	.48	.36	.58	.79	4	10
7. Maladjusted	.18	.06	.28	.46	19	41
9. Scrupulous	.19	.08	.30	.49	17	38
10. Sensitive	.41	.28	.51	.73	5	13
11. Independent	.23	.11	.34	.55	13	30
12. Cheerful	.29	.17	.40	.62	10	22
13. Subjective	.06	−.04	.16	.21	63	141
14. Anxious	.21	.09	.31	.51	16	34
15. Sociable	.29	.16	.40	.62	10	22
17. Warmhearted	.53	.41	.62	.82	4	8
18. Vital	.42	.29	.52	.74	6	13
19. Intolerant	.34	.21	.44	.67	8	17
Md	.29					

The coefficients of the first column in Table 22 express the inter-rater reliability for the ratings of one of the examiners. The median of these coefficients is .29. The second and third columns give the lower and upper 5 per cent significance limits for the coefficients. In the fourth column are given the values for the inter-rater reliability of the pooled ratings from four examiners. The coefficients are computed by the method proposed by Ebel. This method gives for most variables results which are identical with those that would be obtained if we had regarded the examiners as parallel tests and had used Spearman-Brown's formula to estimate the reliability of the four sets of ratings. For a few variables there would be very small differences.

If we regard the data from the four examiners as scores from four parallel tests, we can estimate the number of examiners we should need in order to obtain inter-rater reliability coefficients of a certain level. In the last columns of Table 22 are given the number of times we had to lengthen the test—the number of examiners we had to use—in order to obtain inter-rater reliability coefficients of .80 and .90.

C. Comments

Two main facts have been pointed out in the general discussion. First, the only source of error which can influence our ratings is lack of objectivity. Second, this lack of objectivity is to a high degree dependent on the instruction given to the examiners. With a sufficiently careful and exact instruction, which covers every possibility of evaluating, and gives very little interpretation, it is possible to reach a very high level of agreement. In our case the examiners have had no guidance by such an instruction, as they were left free to use any method they found useful. They have also worked quite independently of each other. From this point of view the inter-rater reliability coefficients obtained are *minimum* values of the objectivity of the method.

However, the only source of error in this case is probably lack of objectivity. This means that the deviation from perfect reliability might have been still greater if other sources of error had been included. From this point of view the results are not promising. The levels of the coefficients are generally so low that, for ability tests, they should definitely be considered as showing the tests to be useless. In Chapter 9, however, we will show that the results are not merely unsatisfactory; they may be used for increasing the usefulness of the method.

Chapter 8. Validity

1. SOME EARLIER INVESTIGATIONS

a) *Sanford et.al.* In an intensive study Sanford investigated the relation between ratings of observable manifestations of thirty-six needs and ratings based on the TAT for the same needs (Sanford et.al., 1943). The subjects, who ranged from 5 to 14 years of age, consisted of twenty-one boys and twenty-seven girls. The ratings of manifest behavior were the combined judgments of three psychologists and the children's teachers. The TAT was administered to each child three times at intervals of about one year. The task for the test examiners was to estimate on a five-point scale the frequency and intensity with which various needs, presses and outcomes were present in a story or a set of stories. The correlation coefficients for the relation between the ratings of overt and covert needs were estimates of the product moment coefficients based on the grouping of subjects in the extremes of the distributions for the two variables (see Kelly, 1939, pp. 17–24). Of the coefficients twenty-seven were positive, four were zero and five negative. Coefficients of .30 or more were obtained for the following needs: Nurturance .41, Construction .37, Retention .34, Harmavoidance .33, and Acquisition, personal .30 (Sanford, 1943, p. 283).

b) *Hartman.* Hartman studied the relation between ratings from the TAT and criterion ratings, based on tests, interviews and case histories, performed independently by the experimenter and a psychiatrist on thirty-seven psychiatric-clinic boys of average intelligence and between the ages of 13–17 (Hartman, 1949). The ratings were performed on a five-point scale for thirty-four personality variables, and the correlation was estimated by biserial coefficients. Of the thirty-four coefficients for the correlation between the TAT ratings and the ratings performed by the experimenter, twenty-seven were positive and seven negative. The mean

correlation was .15. For the correlation between the TAT ratings and the ratings by the psychiatrist twenty-five positive and nine negative coefficients were obtained. The mean correlation was here .17. The results for those four variables, which are closely similar to four of the variables in the present investigation, are reported in the table below.

Personality variable	TAT/Exp.	TAT/Psych.
Dull — Bright	.39	.30
Energetic — Apathetic	.23	.32
Depressed — Happy	.22	.28
Self-reliant — Dependent	.29	.32

Hartman concludes that the results indicate on the whole a constant and direct relationship between response on the TAT and overt behavior.

c) *Harrison.* For intelligence, rated through blind analysis of thirty-seven cases, Harrison obtained a correlation coefficient of .78, when IQs were used as criteria (Harrison, 1940, pp. 128–38).

d) *Davids and Rosenblatt.* In a series of recent research a personality syndrome has been defined and assigned the term the "alienation syndrome" (Davids, 1955, pp. 21–27 and 1956, p. 330). The syndrome is defined by relatively high standing on the negative personality dispositions of pessimism, distrust, egocentricity, anxiety and resentment and relatively low standing on the positive dispositions of optimism, trust and sociocentricity. For a study of the possibility of predicting this syndrome from the TAT two additional positive variables were added, so as to give an equal number of positive and negative variables: Engagement, defined as "an ability to persist at a task, to be seriously and competently engaged in it, to master anxiety and carry through, to control anxiety once it is recognized and to master the state or situation", and acceptance (Davids and Rosenblatt, 1958, pp. 145–52). The positive direction of the syndrome is termed "affiliation". Each of the positive variables corresponds to one of the negative ones. Rank orders of twenty male undergraduates for this syndrome, based on the pooled ratings from the TAT in the ten

variables correlated .44 with a rank order for the syndrome, made up by Murray and based on clinical information, gathered by him for the individuals in a three-year study. For twenty other undergraduates the rank correlation coefficient was .40. If a rank order for the syndrome, based on the students' responses to a variety of direct and projective measures was used as a criterion, the validity coefficient was .67 for the second group.

2. THE PRESENT INVESTIGATION: VALIDITY OF THE EXAMINERS' RATINGS WITH THE STUDENTS' RATINGS AS A CRITERION

A. Statistical procedure and results

The coefficients for the correlation between the examiners' ratings, based on the TAT, and the students' ratings were computed for one examiner at a time over all individuals on the transformed scores. The degrees of freedom = 57 (Lindquist, 1940, pp. 221–2).

The coefficients for the correlation between the ratings by the test examiners and the students are given in Table 23 for each examiner together with the mean validity coefficients. The coefficients are corrected for attenuation in the criterion. The correction has only little effect on the size of the coefficients, since the inter-rater reliability coefficients for the students' ratings are high (pp. 32–33).

If we first look at the coefficient for one variable at a time, there are only two variables for which the mean validity coefficients are significant at the 5 per cent level. They are variables 10, Sensitive (.28) and 17, Warmhearted (.26). Near the 5 per cent level of significance is variable 12, Cheerful (.24).

There is no variable for which all the four examiners show significant validity coefficients. The best one is the variable 17, Warmhearted, for which three of the examiners have significant coefficients at the 10 per cent level at least.

If we look at the columns of the table, there are marked differences between the examiners in the mean level of the coefficients and in the number of positive and in the number of significant coefficients. Examiner B has no negative validity coefficient, but

TABLE 23. *Validity coefficients (corrected for attenuation in the criterion)*

Variable	A	p	B	p	C	p	D	p	$\bar{r}$
1. Extrapunitive	–.12		.05		.14		.01		.02
2. Ostentatious	.12		.20		.11		.07		.12
3. Dominant	.00		.02		.18		.13		.08
4. Oppositional	.14		.00		.18		.11		.11
6. Acute	–.04		.32	<.05	.20		.24	<.10	.18
7. Maladjusted	–.16		.23	<.10	.17		.12		.09
9. Scrupulous	.10		.20		–.15		–.01		.03
10. Sensitive	.22	<.10	.33	<.01	.09		.50	<.001	.28
11. Independent	–.03		.32	<.05	.11		.20		.15
12. Cheerful	.10		.34	<.01	.15		.39	<.01	.24
13. Subjective	–.24	<.10	.10		–.11		–.04		–.07
14. Anxious	–.25	<.10	.24	<.10	.16		.27	<.05	.11
15. Sociable	.16		.12		.33	<.01	.29	<.05	.22
17. Warmhearted	.24	<.10	.33	<.01	.18		.27	<.05	.26
18. Vital	–.15		.19		–.05		.12		.03
19. Intolerant	–.03		.37	<.01	.04		.23	<.10	.15

N = 63 Df = 57

six positive coefficients which are significant at the 5 per cent level at least. On the other hand, rater A has only seven positive coefficients of which none is significant at the 5 per cent level. (The differences between the examiners and the problem as to whether there exists a true difference in their ability to predict the students' ratings, are dealt with in Chapter 10.)

The significance levels of the validity coefficients, given in the probability values for each coefficient, are based on two-tail tests of significance. This test is relevant, when we test a null hypothesis about the relationship between the students' and the examiners' ratings for one variable only. It gives the probability for a coefficient of that size, either positive or negative, by chance. If we had tested a hypothesis of a positive direction of the coefficients the p-values would have been half of the values reported.

However, if we consider only the single validity coefficients, we lose one important aspect of the validity of the test: the question of the whole set of validity coefficients in one column of Table 23. In Figure 4 the distributions of the validity coefficients

for each column are given. The problem may be stated for each one of the distributions: Is the distribution of validity coefficients for an examiner different from that expected by chance? Our question now deals with the set as a whole, not with the individual member of the set.

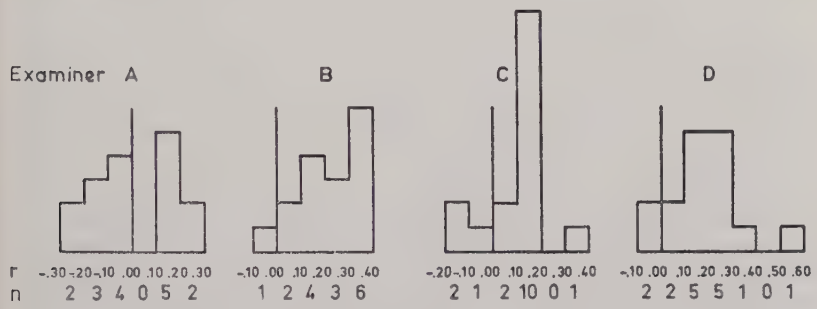


Fig. 4. Distribution of validity coefficients for various test examiners.

There are statistics for testing the significance of an entire set of experimental results (see e.g. Jones, 1953, pp. 375–82). The most exact of them, however, assume independence of the single results. As the validity coefficients of each distribution in our case are not independent, it is more or less meaningless to apply a refined exact method for the study of the problem. Therefore we have chosen a crude sign test for our question, based upon the binomial distribution (see Guilford, 1956, p. 249). The test is applied as a one-tail test for the hypothesis that the distribution of coefficients tends to be higher than that expected by chance.

The sign test shows that the probability for fifteen or more positive coefficients (Examiner B) by chance is about .0003, for fourteen or more positive coefficients (Examiner D) by chance about .002 and for thirteen or more positive coefficients (Examiner C) by chance about .01. The distribution of validity coefficients for examiner A is seen directly to be attributable to chance.

The sign test also assumes independent values, and ours are not independent, because the variables are correlated for both students and examiners. That means that the p-values for the significance levels must be interpreted with caution. The significance level, at least for examiners B and D, and for the mean validity coeffi-

cients are, however, so high that they indicate a real difference between the obtained distributions and the distribution we could expect by chance.

B. Comments

It has been earlier pointed out that some characteristics of the students' ratings must be borne in mind, when the value of the validity coefficients is being interpreted. Some other points of view, which can be relevant for the interpretation will be discussed here.

a) *Blind analysis.* The judging of the test protocols has been performed as blind analysis. This is necessary in scientific connections in order to control some irrelevant factors. In practice blind analysis is the procedure followed in judging protocols from group administered test sessions. It means that for such data the validity coefficients obtained by blind analysis are a relevant expression of the validity of the ratings. In clinical work, however, the judging of a protocol is seldom performed by blind analysis. In such situations the observation of the subject is important for the right interpretation of the projective material in the protocol. Not only direct observations of behavior, which can be performed in every situation, e.g. in an interview, but the behavior in relation to the projective test material is usually expected to be useful for the test examiner. This means that the validity coefficients, calculated for data obtained by blind analysis, will express the minimum validity of the method. Higher validity coefficients would probably have been the result if this other information had been available to the test examiners.

b) *Differences in content of the variables.* In the introduction, the self-evident fact that the validity of the test is investigated for the aspects, characterized by the criterion was pointed out (p. 18). Some characteristics of the students' ratings as criteria have been mentioned in various connections earlier. One aspect will be discussed here. The definitions of the variables were formulated as clearly as possible, in order to minimize interpretation of behavior in the students' ratings. However, in this respect there are great differences between the definitions for various variables. Now the important question is whether the students and the test

examiners interpret the definitions in the same way and give the variables the same content. This is probably, among other things, due to the amount of interpretation between the observed behavior and the rating. One important factor in the interpretation of the variables is the value loading, which the variable has for the interpreter. If a variable has different value for the interpreters, their comprehension of the definition of the variable will be influenced. The schools that have taken part in the investigation are different in character. Probably there exist special value systems in some schools that differ from those of the others. The variables Extrapunitive, Scrupulous or Intolerant, for example, are probably evaluated and interpreted in different ways in a religious school, with strict norms of the behavior, than in another, where the school regulations are very liberal. One main difficulty of the test examiners in predicting the ratings by the students may have been this circumstance.

c) *The rating of all subjects in all variables.* In the present investigation the test examiners were expected to rate all subjects in all variables. This is not the case in clinical work. The test examiners themselves decide which variables they can rate in each case. In this investigation the test examiners also had the possibility to express the subjective confidence in the ratings. This problem is however complicated by the fact that most investigations into the relation between subjective confidence—which for the clinical psychologist determines his choice of variables to rate —and actual ability have failed to show a systematic positive correlation (see Chap. 11).

d) *Mean validity.* In accounts of validity investigations with projective methods, where predictions have been performed in various variables, the results of the investigations are often evaluated from the mean validity coefficients. A low mean validity coefficient for a set of variables is taken as an indicator of low validity throughout. From many points of view this clearly does not do justice to the methods. First, often more variables than can possibly be rated are included in investigations of explorative character. As an example the test examiners in the present investigation were uncertain about the possibility of judging some of the variables and expressed their uncertainty before the begin-

ning of the study. These variables were, however, included in order to study the possibility. And second, it is unreasonable to claim that it is possible to give a broad valid description of a range of individuals on the basis of the single sample of verbal behavior that a TAT-protocol is. There has been a tendency by some projective testers to rely too much on their methods, and by some critics to reject the test if this claim is not fulfilled. The meaningful work of the future will be to collect more scientific data for the possibility to judge personality variables from the TAT and on the basis of these data determine whether some variables can be consistently judged with validity and which these variables are. From this point of view the data obtained in this investigation may be considered as promising.

3. VALIDITY OF RATINGS IN THE "ALIENATION SYNDROME"

A. Problem

In the introduction to this chapter we accounted for an investigation performed by Davids and Rosenblatt with TAT-ratings on some variables forming the alienation syndrome (p. 88). In the present investigation the task for the examiners was not to predict the variables of this syndrome. Neither was the special, relatively objective scoring technique, used by Davids and Rosenblatt, followed by the test examiners. However, some of the variables defined in this investigation correspond rather closely to some of the variables in the syndrome. Those of our variables corresponding most closely to those in the alienation syndrome are, 7, Well-adjusted—Maladjusted, 12, Cheerful—Gloomy and 14, Confident—Anxious. The last two variables directly correspond to two pairs of variables in the alienation syndrome and the first, variable 7, may be said to cover the main content of the definition of the syndrome. It will also be remembered, that these variables were the ones which both examiners and students agreed should be included in the same cluster (pp. 70–71). For this reason, we have studied the validity of the test examiners' ratings in these three variables together.

B. Procedure and result

The pooled ratings in standard scores from the three variables have been used as the rating made by each examiner for each individual on the syndrome. As a criterion of the ratings for the syndrome the pooled ratings of the standard scores for the students' ratings were calculated.

The validity of the pooled ratings for the alienation syndrome was estimated by product-moment correlation for examiners B, C and D. For examiner A no coefficient was calculated since two of the single variables show negative validity coefficients and only one a positive coefficient. The validity coefficients obtained were .24 (Examiner B), .19 (Examiner C) and .36 (Examiner D). For the correction for attenuation in the criterion, the reliability of the pooled ratings by the students in the three variables was estimated by means of the formula for the reliability of test batteries, given by Ekman (1947, p. 36). The coefficient for this reliability = .85. In Table 24 are given the validity coefficients for the alienation syndrome, with validity coefficients corrected for attenuation in the criterion.

TABLE 24. *Validity coefficients for pooled ratings in the alienation syndrome*

Examiner	B	C	D
r_c	.26	.21	.39

N = 63 Df = 57

C. Comments

The result is what might be expected from the validity coefficients of the ratings in the single variables. For two of the examiners the coefficients are significant, for rater B at the 5 per cent level and for rater D at the 1 per cent level. In the interpretation of the results we must remember the facts to which we referred in the introduction to this section. However, the result supports the findings of Davids and Rosenblatt, that it is possible successfully to rate the syndrome from the TAT. This is important, since knowledge of an individual's relative standing in this syndrome may be of practical importance, e.g. in a personnel selection situation.

4. THE RELATION BETWEEN TAT-RATINGS OF LATENT BEHAVIOR AND THE STUDENTS' RATINGS

A. Problem

The ratings of latent behavior by examiner C have not been performed as predictions of the students' ratings. It is, however, interesting to compare the coefficients for the correlation between C's predictions of manifest behavior and the students' ratings with the coefficients for the correlation between the same examiner's ratings of latent behavior and the students' ratings. As the agreement between the ratings of latent behavior and the predictions of the students' ratings was very high for most variables, we cannot expect great differences in the size of the coefficients in the computations here. On the basis of the well-known theory that for some personality variables the latent tendencies are suppressed and will not be expressed in manifest behavior, we might however assume that the coefficients for the correlation between the test examiner's ratings of manifest behavior and the students' ratings should be higher than the coefficients for the correlation between the examiner's ratings of latent behavior and the students' ratings.

B. Result

The coefficients for the correlation between examiner C's ratings of latent behavior (Cl) and the students' ratings are reported in Table 25. For comparison are also shown the coefficients for the correlation between the predictions of the students' ratings (Cm) and the criterion. The coefficients in both the columns have been corrected for attenuation in the criterion.

I we compare the two columns, we find that the ratings of latent behavior give as many positive correlations as do the predictions of the students' ratings. In half of the variables the ratings of latent behavior show higher agreement with the students' ratings than do the predictions of the students' ratings. There is no systematic difference in the direction or the size of the coefficients in the two columns. The result does not support our assumption of higher coefficients for the correlation between the predictions of the students' ratings and the students' ratings than for the

TABLE 25. *Coefficients for the correlation between the students' ratings and ratings of latent (Cl) and manifest (Cm) behavior for examiner C*

Variable	r_{Cl}	p	r_{Cm}	p
1. Extrapunitive	.03		.14	
2. Ostentatious	.14		.11	
3. Dominant	.20		.18	
4. Oppositional	.06		.18	
6. Acute	.20		.20	
7. Maladjusted	.21		.17	
9. Scrupulous	−.20		−.15	
10. Sensitive	.05		.09	
11. Independent	.08		.11	
12. Cheerful	.20		.15	
13. Subjective	−.08		−.11	
14. Anxious	.38	<.01	.16	
15. Sociable	.23	<.10	.33	<.01
17. Warmhearted	.07		.18	
18. Vital	−.06		−.05	
19. Intolerant	.08		.04	
M	.10		.11	

correlation between the ratings of latent tendencies and the students' ratings.

5. TAT-RATINGS AND OBJECTIVE TEST DATA

In the result of the group test administered we have an objective criterion of the main aspects of variable 6, Acute (p. 28). The students' positions in the distribution of the test scores were determined by the pooled standard scores on the two sub-tests. For the estimation of the validity of the TAT-ratings in variable 6, the coefficients between the pooled scores and the TAT-ratings were computed for each test examiner. The calculations were performed for one group at a time. The obtained coefficients were transformed to one coefficient for each examiner by means of Fisher's z. In Table 26 are reported the coefficients corrected for attenuation in the criterion.

All the coefficients are positive. The coefficient for examiner B is significant at the 1 per cent level and the coefficient for exam-

TABLE 26. *Coefficients for the correlation between TAT-ratings in variable 6, Acute, and test scores*

Examiner	A	B	C	D	M
r_c	.18	.48	.31	.17	.29

Df = 46

iner C is significant at the 5 per cent level. The coefficient for rater B is of considerable size. The coefficients must be evaluated in the light of the fact that the instruction to the examiners was not to predict the group test data but to predict the students' ratings of each other in the variable Acute—Dull.

6. VALIDITY AND AGREEMENT IN STRUCTURE

A. Procedure and results

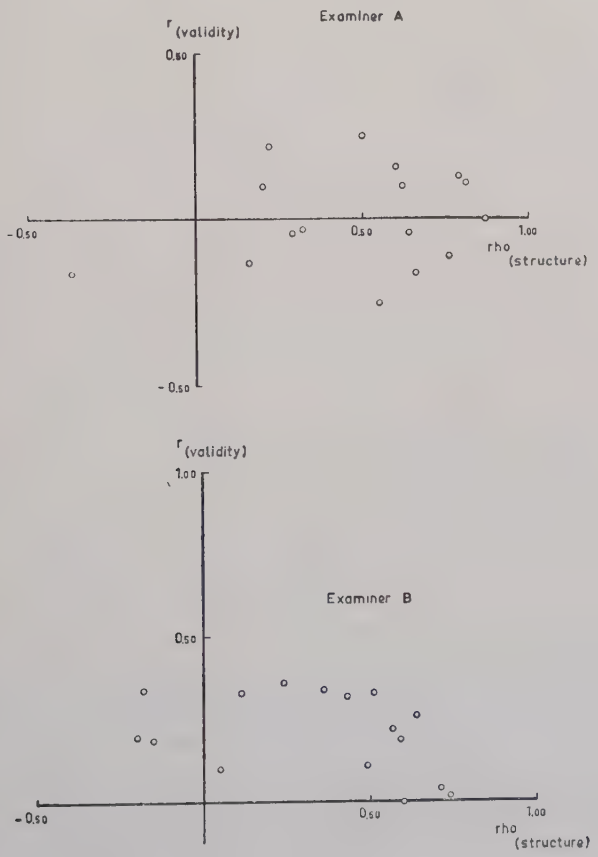
For each examiner rank-order coefficients have been calculated for each variable, expressing the degree of agreement between the rank order of the intercorrelation coefficients in each column in the correlation matrix for the examiner's ratings on the one hand, and the rank order of the intercorrelation coefficients in the same column of the matrix for the students' ratings on the other. For each examiner we thus obtained sixteen coefficients, one for each variable (pp. 71–72). The mean of these coefficients for each examiner was found to be:

Examiner	A	B	C	D
rho	.49	.34	.45	.47

a) *Problem I.* We can now relate the size of the validity coefficients for various examiners to the degree of agreement in structure, as this is expressed in the coefficients reported above. The mean coefficients for three examiners (A, C, D) are of about the same size. One examiner differs markedly from the others in his mean coefficient for agreement in structure with the students. It is examiner B. He has obtained a mean coefficient, which indicates the lowest degree of agreement in structure with the stu-

rents' ratings. He is also the examiner, whose ratings show the greatest agreement with the students' ratings, when we compute validity coefficients over all individuals for one variable at a time. On the one hand, examiner B thus shows the least agreement in structure with the students' ratings, but on the other hand, his single ratings show the greatest agreement with the students'.

b) *Problem II.* The above comparison applied to the relation between the mean levels of agreement as regards structure and the mean levels of the size of the validity coefficients. Yet the possibility exists that there is a positive relation for each examiner between the degree of agreement in structure and the size of the validity coefficient for each variable. The easiest way of studying



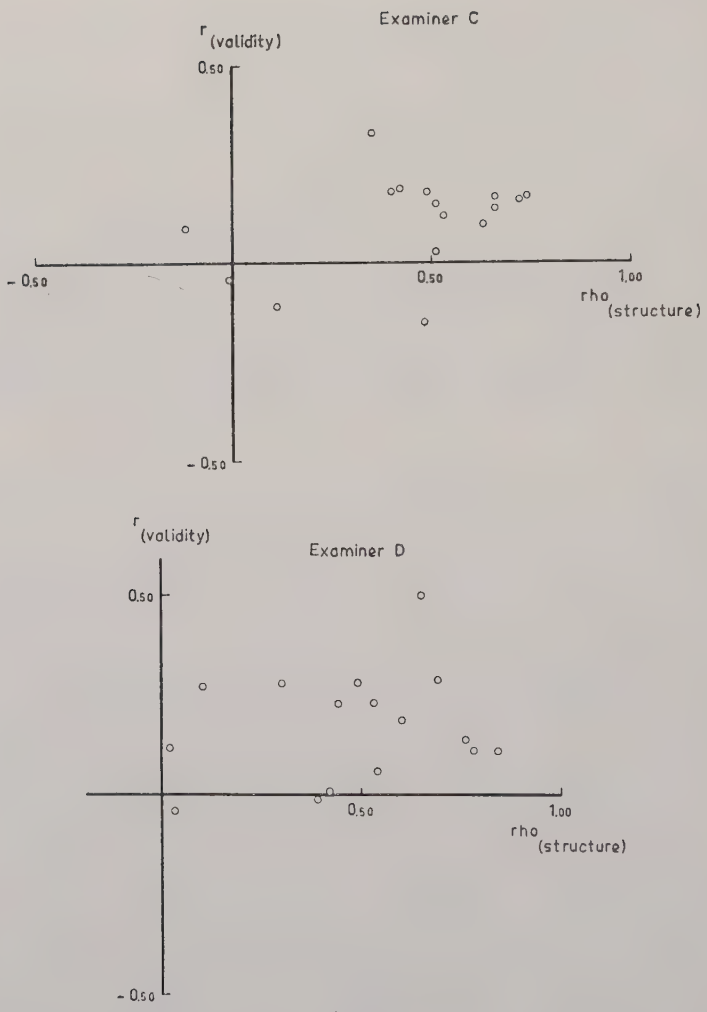


Fig. 5.

this problem is by plotting. The plots are given for each examiner in Figure 5. The values for agreement as regards structure (along the χ-axes in Fig. 5) for each variable are the ones reported in Table 17 (p. 71).

From the plots we can find no common tendency for all the examiners. For examiner B we find a tendency to a reversed rela-

tion between the degree of agreement with the students in structure and the size of agreement in ratings of the individuals.

B. Comments

On the basis of the results given to illustrate the problem, we must conclude that our data do not indicate any systematic relationship between the degree of agreement in structure between examiners' and students' ratings on the one hand, and the size of the validity coefficients on the other. The two seem to be independent of each other.

Chapter 9. Reliability and validity

1. GENERAL DISCUSSION

On the basis of the reported results, we will devote this chapter to a discussion of the relation of reliability to validity, and some consequences of this relation.

We have found that there exist very few inter-rater reliability coefficients of considerable size (p. 85). With this result in mind it is interesting to study the validity coefficients in Table 23 (p. 90). We are in the situation—very unusual in test construction—of having validity coefficients of about the same size as the reliability coefficients.

This situation is unusual but not unfavourable. We have test examiners, each of whom, while employing different methods, makes use of relevant, though not necessarily identical information. How these findings may be applied in future research will be discussed in a later connection (see chapter 19). In this chapter a practical consequence will be discussed and illustrated.

2. A PRACTICAL CONSEQUENCE

A practical consequence, which may have relevance also to other projective methods, is possible to realize immediately. This is to combine ratings from more than one test examiner without giving detailed instructions concerning the judgments. By combining very few test examiners' ratings it is possible to attain good validity coefficients, if every additional examiner extracts and contributes new relevant information, which was not used by the earlier ones.

A. The present investigation: Procedure and results

In the present case, we have already the validity coefficients for each examiner and the inter-rater reliability coefficients for each pair of examiners. By using the validity coefficients and the inter-

rater reliability coefficients for each variable we can now compute the mean validity coefficients for the pooled ratings of two examiners. The mean validity coefficients, found by the formula given by Peters and van Voorhis (1940, p. 196) are reported in Table 27, the first column. The computations are based on validity coefficients, corrected for attenuation in the criterion.

In Chapter 10 will be discussed whether there exists a true rank order between the examiners as regards their ability to predict the students' ratings. We have found that the results support such a hypothesis. The results indicate that there is only a slight possibility that the rank order, found between the examiners, could be obtained by chance. Therefore it may be reasonable to compute the validity coefficients for the pooled ratings received from examiners on various levels of ability. Thus in Table 27 are also reported the validity coefficients for the pooled ratings made by two of the three examiners B, C and D, and the validity coefficients for the pooled ratings of examiners B and D, whose coefficients indicate the highest level of rating ability.

TABLE 27. *Validity coefficients for pooled ratings by two examiners*

Variable	Two ex. of A, B, C, D		Two ex. of B, C, D		Ex. B, D	
	r	p	r	p	r	p
1. Extrapunitive	.03		.08		.03	
2. Ostentatious	.15		.14		.16	
3. Dominant	.12		.15		.10	
4. Oppositional	.12		.11		.07	
6. Acute	.21		.30	<.05	.34	<.01
7. Maladjusted	.10		.22	<.10	.21	
9. Scrupulous	.05		.00		.13	
10. Sensitive	.36	<.01	.39	<.01	.52	<.001
11. Independent	.19		.27	<.05	.33	<.01
12. Cheerful	.31	<.05	.37	<.01	.47	<.001
13. Subjective	−.10		−.02		.04	
14. Anxious	.14		.33	<.01	.37	<.01
15. Sociable	.29	<.05	.27	<.05	.24	<.10
17. Warmhearted	.30	<.05	.31	<.05	.36	<.01
18. Vital	.04		.10		.17	
19. Intolerant	.18		.27	<.05	.39	<.01

B. Comments

The combination of ratings from two examiners in our case leads to coefficients of considerable size for some variables, especially when the test examiners combined are examiners B and D. When the protocols are once collected the profit in validity of the ratings may be worth the expenditure of time and money to add one or more test examiners to the first one. Besides, there are investigations which indicate that, for some purposes, protocols from group administrations may yield as much relevant information as protocols from individual test sessions (Eron and Ritter, 1951, pp. 55–61. Lindzey and Heinemann, 1955, pp. 34–55). This is especially the case in other than clinical situations. In summarizing the results of their investigation Lindzey and Heinemann write:

> "In view of the many studies that have been executed using TAT-procedures it is reassuring to find so little evidence indicating important differences between the group and individual tests. The possibility of using the group test with confidence, even under limited conditions, means that the empirical development of the TAT can proceed at an infinitely faster rate and that many problems, as well as applications, that hitherto have been unapproachable can now be dealt with." (p. 54).

By group administration it is possible to collect data from several individuals in one test session. The time saved may be used to combine ratings from more than one test examiner. With group administration it is also possible to collect samples of verbal behavior in a relatively short time from more than one occasion, which would increase the amount of relevant and necessary information for valid ratings. Distribution over several occasions may be important also with reference to results, which indicate that the TAT stories are influenced by momentary, situation-bound factors. Perhaps the total test time may be better used in some situations when distributed over more than one occasion with a small number of pictures in group administrations each time.

Chapter 10.
Rank order between test examiners and variables

1. RANK ORDER BETWEEN TEST EXAMINERS

A. Problem

From Table 23 (p. 90), where the validity coefficients are reported for each test examiner, we can by inspection find a marked difference between the test examiners in the mean levels of the coefficients. If we consider the number of significant coefficients we obtain the order B, D, C and A. The mean coefficients for the test examiners give the same order. In Fig. 6 is shown a clear picture of the tendency. It shows how many times each of the test examiners has obtained the first, the second, the third and the fourth rank, if the validity coefficients are ranked in order of size for each variable.

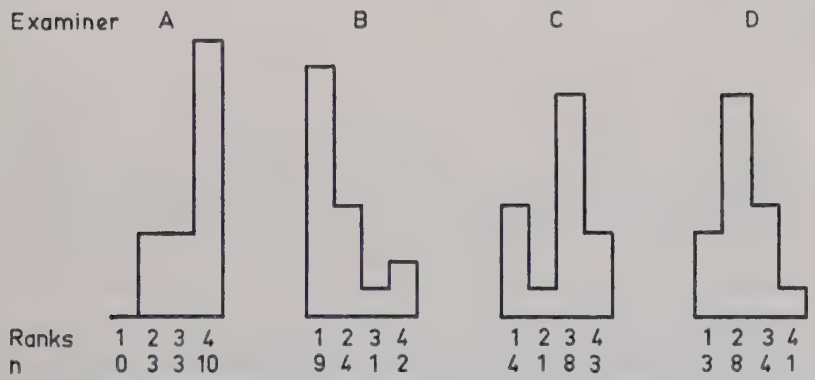


Fig. 6. Distribution of ranks for various test examiners, when the validity coefficients for each variable are ranked in order of size.

The problem may now be stated as follows: Does there exist a true rank order between the examiners in their ability to predict the students' ratings? How great is the probability that the rank order formed may be obtained by chance?

B. Statistical procedure

Our problem is difficult to study by exact, statistical methods on account of the non-independency of the coefficients. With this in view the following crude method has been used. As mentioned above, the validity coefficients were ranked in order of size for each variable. This rank order may be looked upon as a rating of the ability of the examiners to predict the students' ratings for each of the variables. The rank orders are shown in Table 28.

TABLE 28. *Rank orders of validity coefficients for each variable*

Variable	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19	M
Examiner																	
A	4	2	4	2	4	4	2	3	4	4	4	4	3	3	4	4	3.438
B	2	1	3	4	1	1	1	2	1	2	1	2	4	1	1	1	1.750
C	1	3	1	1	3	2	4	4	3	3	3	3	1	4	3	3	2.625
D	3	4	2	3	2	3	3	1	2	1	2	1	2	2	2	2	2.188

The degree of agreement between the rank orders was estimated by Kendall's coefficient of concordance (Edwards, 1954, pp. 402–7).

C. Result

The computations give a coefficient of concordance of .311 for the agreement between the rank orders in Table 28. As N is here only four, the significance level of the coefficient was determined according to Siegel (1956, p. 235–6). An s-value of 398.0 was obtained. This is well above the value required for significance at the 1 per cent level (s = 288.7 Siegel, 1956, p. 286).

The average intercorrelation between the rank orders for the test examiners calculated from the coefficient of concordance is .265. This value gives the mean correlation between any two rank orders. From the mean intercorrelation a reliability coefficient of the average rank order that is based on all the sixteen rank orders was calculated. The coefficient was .852. This value indicates the reliability in the final rank order B, D, C, A.

The method for the computation of a coefficient of concordance assumes independent ranks. As the ratings by the students are

correlated to some extent, the validity coefficients are not independent. Therefore the s-value cannot be used as an exact value for the estimation of the significance level for the final rank order between the examiners. However, the dependency may be judged as small in our case. The result may therefore be considered to support the hypothesis that there exists a true rank order between the examiners in their ability to predict the students' ratings.

D. Comments

The significance level of the coefficient of concordance indicates that the probability of obtaining this set of rank orders between the examiners in their ability to judge the TAT protocols by chance is small, when the task is that which is given in this investigation. This means that a positive result with the TAT is dependent on which test examiner interprets the test stories. The rank order may be due to a) differences in rating ability of various test examiners, b) differences in the usefulness of various methods for the interpretation, or c) a combination of these sources. This problem needs further exploration. Vernon (1957, p. 205) draws the following conclusion from a discussion of the problem:

"... it is hardly possible to dissociate the test from the tester. One clinical psychologist does well with interviewing or with Rorschach, another with Thematic Apperception or drawings, another with deterioration tests or expressive movements and so on. Hence in order to prove the worth of these or other methods, the diagnostician has to rely chiefly on evidence provided by the selector or the experimentalist. Unfortunately he is much too apt to trust to the face validity of case studies and to his own experience; these may be supremely good in individual instances and worthless in others."

2. RANK ORDER BETWEEN VARIABLES

A. Problem

One further practical problem is that of a possible true rank order between the variables in the possibility of judging them from the TAT protocols. If such a rank order exists it may result in agreement between the rank orders of the validity coefficients for the different examiners.

B. Procedure and result

For each examiner the validity coefficients were rank ordered. The rank orders are reported in Table 29. The degree of agreement has been estimated by a coefficient of concordance as was performed for the study of the rank order between examiners.

TABLE 29. *Rank orders of validity coefficients for each examiner*

Variable	A	B	C	D	M
1. Extrapunitive	12	14	9	14	12.25
2. Ostentatious	5	9	11	13	9.50
3. Dominant	8	15	3	9	8.75
4. Oppositional	4	16	5	12	9.25
6. Acute	11	5	2	6	6.00
7. Maladjusted	13	8	6	10	8.75
9. Scrupulous	6	10	16	15	11.75
10. Sensitive	2	3	12	1	4.50
11. Independent	9	6	10	8	8.25
12. Cheerful	7	2	8	2	4.75
13. Subjective	15	13	15	16	14.75
14. Anxious	16	7	7	4	8.50
15. Sociable	3	12	1	3	4.75
17. Warmhearted	1	4	4	5	3.50
18. Vital	14	11	14	11	12.50
19. Intolerant	10	1	13	7	7.75

The coefficient of concordance for the rank orders in Table 29 is .465. The corresponding χ^2-value is 27.860, which is well above the level for significance at the 5 per cent level ($\chi^2 = 24.996$). The size of χ^2 indicates that there is substantial agreement. The reliability of the final rank order between the variables is expressed in a coefficient of .616.

Attention may here be drawn to the fact that the rank order between the variables is affected by the size of the validity coefficients. To the same degree as the validity coefficients for a certain examiner have been obtained by chance, to the same degree will this give a rank order which is uncorrelated with those of the other examiners. This fact is illustrated by the circumstance that the highest single rank coefficient between two of

the rank orders in Table 29 is that between examiners B and D (rho = .63 $p < .01$), who have the greatest validity.

We must judge the result in the light of the fact that our validity coefficients are not independent, as the exact use of the coefficient of concordance assumes. As in the problem of rank order between examiners, the effect of the intercorrelations between the ratings in various variables for students and examiners may be judged to be small. If the effect were of important size it would appear in groupings of coefficients for those variables which show high intercorrelations. There are, however, no such systematic groupings.

C. Comments

As mentioned above, the χ^2-value cannot be used as an exact estimate of the significance level for the agreement between the examiners in their rank orders for validity coefficients. The result however supports the hypothesis that there exists a true rank order between the variables, so that the ratings by the students are predicted with greater accuracy for some variables than for others.

The variables with the lowest mean ranks (predicted with high validity) are the the following: 17, Warmhearted (M = 3.5), 10, Sensitive (M = 4.5), 12, Cheerful and 15, Sociable (M = 4.75). For these four variables the mean validity coefficients are higher than .20 (p. 90). If we only take into consideration the rank orders for examiners B and D (see discussion above) the following variables have the lowest mean ranks: 10, Sensitive and 12, Cheerful (M = 2.0), 19, Intolerant (M = 4.0) and 17, Warmhearted (M = 4.5). The mean rank for variable 15, Sociable, is here higher on account of the low validity coefficient for examiner B.

The variables with the highest mean ranks (predicted with low validity) are the following: 13, Subjective (M = 14.75), 18, Vital (M = 12.5), 1, Extrapunitive (M = 12.25) and 9, Scrupulous (M = 11.75).

Chapter 11. Subjective confidence—actual ability

1. GENERAL DISCUSSION

A. Problem

In usual clinical work the TAT examiner is not expected to rate all test protocols on a fixed number of variables. In many situations he chooses in each case those aspects for which he thinks that he is able to make valid ratings. The question then arises: Does there exist a positive relation between the subjective confidence with which a test examiner performs his ratings and the objective, actual accuracy with which the ratings are given? The problem of the validity of subjectively experienced confidence is important also in other situations than that mentioned.

The general problem of subjective confidence in ratings has two aspects. The first is concerned with the relation between subjective confidence and actual ability *over individuals* in one variable. If there exists such a positive relation, a subject who has been rated with great subjective confidence in a certain variable has also actually been predicted with greater objective accuracy.

The second aspect is the problem of the relation between subjective confidence and actual ability *over variables.* If there exists such a positive relation it means that the higher the level of subjective confidence is for variables the higher are the validity coefficients.

The two problems are statistically independent of each other. The first problem relates the level of subjective confidence over individuals to the differences between the experts' ratings and the students' ratings in one variable at a time. The second problem relates the level of subjective confidence over variables to the level of the validity coefficients. It is possible to obtain a positive relation for the first problem and a negative relation for the second or vice versa. Thus the two problems must be studied with different methods.

B. Some earlier findings

a) Bruner and Tagiuri (1954, p. 642) summarize a discussion of the relationship of confidence to accuracy by saying: "Some writers find a positive relationship (e.g. Odbert, 1934), some a negative one (e.g. Steinwutz, 1947) and some none (e.g. Polansky, 1941)." The interpretation of the results as ambiguous may be due to the neglect of the fact that the problem has two aspects. As long as we do not know the experimental details, we do not know for which problem the results are relevant.

b) In a great program for predicting the success of candidates in clinical psychology by means of a great many objective test data and rating procedures, Kelly and Fiske tried to estimate the relation between subjective confidence and objective accuracy as it is formulated in our second problem (Kelly and Fiske, 1950, p. 406). They used the standard deviations of the experts' ratings as an operational definition of the subjective confidence in the ratings. The result was that with increasing confidence—increasing standard deviations of the ratings—there was a tendency to lower validity coefficients. In a discussion of projective methods in another connection, Rotter (1953, p. 297) summarizes his impressions of the investigation by Kelly and Fiske:

> "An overview of this study leaves the reviewer with two general impressions. One of them is the striking lack of relationship between the experts' confidence in themselves and their methods (even in situations and problems where they had never tried these methods before) and their actual ability to predict. It reemphasizes the inadequacy of the practice of approaching test validity through 'clinical experience' or 'expert' pronouncement."

2. THE PRESENT INVESTIGATION: THE RELATION BETWEEN SUBJECTIVE CONFIDENCE AND OBJECTIVE ACCURACY OVER INDIVIDUALS

A. Problems

In this section we have studied the first of the problems discussed in the first part of the chapter—the relation between the subjective confidence with which an examiner has performed a rating of an individual and the objective accuracy with which the rating has been made. For the study of this problem we need two

measures; one for subjective confidence and one for actual ability of rating. We will study the two measures before relating them to each other.

B. Individuals, rated with great or little subjective confidence

a) *Problem.* Each time an examiner delivered a rating of a subject, he also rated the subjective confidence, with which he performed the prediction of the students' ratings (p. 59). Now we can put the question: Do there exist for each rater some individuals, who throughout—over variables—are rated with greater subjective confidence than others? If this can be shown to be the case, it is an interesting fact in itself. Besides, we can use one measure throughout for each subject, to study the relation between subjective confidence for individuals and other variables.

For each rating of a subject we have obtained one rating of the subjective confidence with which the rating was performed. The sum of these ratings made by each examiner may be used as an expression for the subjective confidence with which the examiner has rated a certain subject. Our problem then is: How reliable is this measure of subjective confidence in rating individuals?

b) *Procedure and results.* The reliability has been estimated for each examiner over all individuals by means of Kuder–Richardson's formula (Guilford, 1956, p. 454). The reliability coefficients, estimated in this way are reported in Table 30.

TABLE 30. *Reliability of pooled ratings of subjective confidence*

Examiner	A	B	C	D
r_{tt}	.62	.72	.81	.69

It will be remembered that examiner B did not perform any ratings of subjective confidence for one group consisting of seventeen subjects. Therefore the reliability coefficient for his ratings are obtained for 46 subjects, while the reliability coefficients for the other examiners are based on 63 subjects.

The reliability coefficients may be interpreted from two points of view. First, they indicate that there are some individuals who are throughout rated with higher subjective confidence than

others. There exists a real rank order in this sense between the individuals. Second, the reliability coefficients may be considered high enough to motivate the use of the sum of the subjective confidence ratings by each examiner as an expression of the individuals' positions on the scale of the subjective confidence with which the individuals are judged.

Now that we have found that some individuals are rated with greater subjective confidence than others throughout, the next problem is: How great is the agreement between the examiners about the individuals' positions in the distribution?

For each individual the sum of the confidence ratings given by one examiner was considered as an expression of the degree of subjective confidence with which the individual was rated in relation to other subjects by the examiner. Thus one such value was obtained for each individual and each examiner. An intraclass coefficient was computed according to the method proposed by Ebel for the 46 individuals for whom all the examiners had given ratings of subjective confidence (Ebel, 1951, pp. 402–24). The variance between raters was not included in the error term, since differences in the mean levels did not interest us in this connection. The intraclass coefficient calculated in this way is .618. This value gives the degree of agreement between the examiners in their ratings of subjective confidence. The coefficient can be used as an estimate of the reliability of the pooled ratings of subjective confidence by the four test examiners.

The intraclass coefficient is not very high, when used as an indicator of differential reliability. It indicates, however, that there is a significant agreement between the examiners as to which students are judged with great and which with small subjective confidence. One of the possible sources of this agreement will be investigated later (see p. 121).

C. Individuals, rated with great or little objective accuracy

a) *Problem.* The problem may be stated here in the same way as for subjective confidence: Are there any individuals, who are rated throughout with greater objective accuracy than others? A positive answer to the question has two aspects; it will be an

interesting fact in itself and it will give a measure to use for estimating the relation to other variables, e.g. test length.

b) *General discussion of procedure.* The best measure of the actual accuracy of a test examiner's rating of a particular subject would be the difference between the transformed scores for the test examiner's rating, and the rating made by the students. If, however, the signs of the differences are ignored—which we have to do, as we are not interested in the direction of the differences, only the size—we may introduce a spurious correlation with some other variables, which is due to stereotypes in the ratings. An example is the common tendency to a wider distribution of ratings, when the examiner is confident (see e.g. Ekman, 1951, p. 456). This will lead to greater chance differences for individuals, who are rated with greater subjective confidence. If the signs are ignored and the difference distribution is correlated with the distribution of confidence ratings, we have introduced a spurious correlation. The result will be a positive coefficient for the correlation between differences and subjective confidence ratings, which might be misinterpreted as a negative correlation between subjective confidence and objective accuracy, even if there is no systematic correlation.

The effect, pointed out, must be borne in mind when working with differences without sign, or squared differences, where the role of the sign is also neglected. When the data upon which the calculations are performed are ratings, the effect will e.g. under certain conditions invalidate the use of D^2 (squared differences), proposed by Cronbach and Gleser as a measure for assessing similarities between profiles (Cronbach & Gleser, 1953, pp. 456–73).

c) *Procedure and results.* As we could not use the differences between the ratings made by the test examiner and the students as an expression of the accuracy of the test examiner's rating, the following measure of accuracy was applied. Each subject, for whom a test examiner and the students are in agreement regarding the direction from the mean of the rating in a certain variable, was marked with plus, and each subject for whom the ratings were given different directions was marked with minus. The matrix for the accuracy of the ratings for each test examiner can then be considered as a matrix of test scores, where the result on each test

item for each subject is judged as right or wrong. The sum of all plus signs for an individual can be used as an indicator of the individual's position in the distribution of individuals with respect to the accuracy with which they are rated.

The reliability of this measure of accuracy for individuals was estimated by the formula given by Kuder–Richardson (Guilford, 1956, p. 454). The reliability coefficients are reported in Table 31.

TABLE 31. *Reliability of the measure of objective accuracy for each examiner*

Examiner	A	B	C	D
r_{tt}	.47	.52	.53	.26

N = 63

Used for differential reliability the coefficients are low, especially that for examiner D. They, however, indicate that the distributions of total accuracy scores for the individuals have not been obtained by chance. Some individuals are judged throughout with greater accuracy than others for each examiner.

Our problem now is: How great is the agreement between examiners about the individuals with respect to the accuracy, with which they are judged?

For each individual the plus signs—which for each variable indicate agreement between the test examiner and the students about the direction from the mean of the rating—were summed up for each examiner. Thus we obtained one distribution of such scores in respect of all the subjects for each examiner. Our problem can now be formulated: How do the test examiners agree about the positions of the individuals in these distributions? The problem was studied by means of the intraclass formula proposed by Ebel (1951, p. 410). The variance between raters was excluded from the error term. Thus by the intraclass formula we obtained a coefficient for the reliability in the total distribution of accuracy scores with the variance between raters excluded. In this way, we get an estimate of the agreement on the individuals, which is independent of the general level of ability for various examiners. The intraclass coefficient is .580.

The coefficient is not very high, if used as an estimate of dif-

ferential reliability. It indicates, however, that there are some individuals, who are judged with greater objective accuracy than others throughout.

D. The correlation between subjective confidence and objective accuracy for individuals

a) *Problem.* As we have found in our study of the measures for subjective confidence and objective accuracy that both the measures show significant consistency over individuals for each examiner, we can now relate them to each other.

b) *Procedure and results.* For each examiner the relation between subjective confidence and objective accuracy was estimated by product moment coefficients for the correlation between the distributions of the measures for the two variables. The coefficients are reported in Table 32. In the second column, the coefficients are corrected for deficient reliability in the distributions of accuracy scores (see reliability coefficients in Table 31). In the third column, the coefficients are corrected also for deficient reliability in the distributions of scores for subjective confidence (see reliability coefficients in Table 30).

TABLE 32. *Correlation between subjective confidence and objective accuracy*

Examiner	r	r_c	r_{cc}
A	–.21	–.30	–.39
B	–.19	–.26	–.31
C	.12	.16	.18
D	.22	.43	.52

Half of the coefficients are positive and half negative. The mean coefficient in the first column is –.01. None of these coefficients are significant at the 5 per cent level. As they can be the result of chance factors, the corrected values are given as illustrations rather than as a proof of significant correlation.

There is no systematic relation between the coefficients in Table 32 and the mean validity coefficients for the different test examiners.

Our results indicate no systematic positive relation over indi-

viduals between subjective confidence and objective accuracy, measured, as in the calculations here. The tendency, given in the corrected values in Table 32 for the relation, to significant positive or negative correlation coefficients for various test examiners needs further exploration.

3. THE PRESENT INVESTIGATION: THE RELATION BETWEEN SUBJECTIVE CONFIDENCE AND OBJECTIVE ACCURACY OVER VARIABLES

A. Problem

The second of our problems, as they were formulated in the general discussion in the first part of the chapter, was roughly investigated after each examiner had rated two groups of subjects (Magnusson, 1957, pp. 58–59). The problem is the relation between the level of subjective confidence for ratings in a variable and the actual accuracy with which the variable is judged.

B. Procedure and results

For the estimation of the relation we used the sum of ratings of subjective confidence in each variable as an estimate of the subjective confidence, and the validity coefficient as an estimate of the objective accuracy. Correlation coefficients were calculated for the two groups one at a time and a mean coefficient calculated for each examiner. The coefficients obtained between the mean ratings of subjective confidence for each variable and the validity coefficients were: for rater A .07, rater B .10, rater C .20 and for rater D −.01. None of the coefficients are significant at the 5 per cent level and there is no systematic tendency in the coefficients. Thus we do not find any systematic positive correlation for the relation between subjective confidence and actual ability for variables.

Another illustration of the lack of positive relationship is the result of ratings, obtained beside the original plan (Magnusson, 1957, p. 59). As described (p. 58) four of the variables defined for the study of the five projective methods were excluded from the TAT investigation, since the test examiners were of the opinion that these variables were impossible to rate from the TAT

protocols, if the instructions were to be followed. Through an oversight examiner D, however, rated the subjects of one group (N = 10) also on these four variables (5, Cautious—Impulsive, 8, Negligent—Careful, 16, Reflective—Not reflective, 20, Rational—Emotional). The coefficients for the correlation between these ratings and the ratings performed by the students were .24, .28, .70 and .18 respectively (corrected for attenuation in the criterion). The mean level of the validity coefficients for these variables, considered as impossible to predict, is higher than that for the others (see Table 23 p. 90).

4. CONCLUSIONS

The conclusions and implications of the findings that high subjective confidence is not always accompanied by an objective, actual ability is well expressed by Kelly and Fiske:

"Whatever the explanation may be, our findings would reemphasize the necessity of determining the actual validity of any tool used in a program of personnel selection or in personality diagnosis. And, since most of the techniques employed in the global assessment type of personality evaluations involve the clinician-technique combination, it would appear relevant to ask about the validity of such individual clinician-technique combinations for specific functions.

"Evidence such as that accumulating in this project serves to remind us of the fallibility of the human being both as a measuring device and as an integrator of data. In laboratories, in factories, and in accounting offices, it has been found necessary to supplement his sensory and perceptual capacities with an elaborate array of measuring instruments and computing devices. Pending the gradual development of better measures of psychological variables and comparable aids for combining them, we must continue to rely heavily on human judgment. In so doing, however, we must be continually aware of the magnitude of the errors of such judgments. These errors can be minimized by placing greatest reliance on measures of demonstrated reliability and validity." (Kelly and Fiske, 1950, p. 406.)

The practical consequence for the problem mentioned in the introduction to this chapter (p. 110) might be that it should not be left to the test examiner to choose the variables to be judged from the test protocol in each case. If we can verify the possibility to rate some variables with validity consistently for a certain aspect, these variables should be prescribed for use when this aspect is of current interest.

Chapter 12.
Subjective confidence, objective accuracy and test length

1. SUBJECTIVE CONFIDENCE AND TEST LENGTH

A. General discussion

As most investigations seem to show that there is less positive relationship between subjective confidence and actual ability than we might be justified in expecting, we have studied some aspects of the problem, to which little attention has been paid. The results may throw some light on the important problem of subjective confidence.

It is reasonable to assume that one of the main reasons for the rater's feeling of confidence is the amount of information that he feels that he has at his disposal. An impression of much information as a basis for the rating gives a feeling of confidence when the rating is being made. In the present case the only information the test examiners possessed was the test stories. It is reasonable to hypothesize that the amount of information that the examiners experience is related to the length of the test stories. It is a common attitude among projective testers that the test lenght is of great importance to the usefulness of a test protocol.

B. Procedure and result

The hypothesis that the test examiners' ratings of their confidence is related to the subjectively experienced degree of information, as this is expressed in the length of the test stories, has been tested in the following way.

For each individual the mean value was computed for the confidence ratings which each individual had obtained by each test examiner. For each test examiner and group, rank order coefficients were computed for the correlation between the number of words of the test protocols and the mean confidence ratings. The coefficients were transformed to one coefficient for each examiner

by means of Fisher's z. As we are more interested in the relationship between the two variables correlated than in the accuracy to predict one variable from the other, it is relevant to correct the obtained coefficients for deficient reliability in both the measures.

The reliability coefficients for the measure of subjective confidence have been calculated and reported earlier (p. 112). The reliability in the measure of test length was estimated by split-half technique for the number of words per story. The coefficient for the reliability of total number of words, obtained by means of Spearman–Brown's formula, is .962. An individual who tells long test stories about some pictures will also tell long test stories about the others.

The coefficients (obtained and corrected) for the relation between the test examiners' rating of subjective confidence and test length are reported in Table 33. The p-values reported have been obtained with one tail tests of the significance levels, since the sign of the correlation coefficients was predicted.

TABLE 33 *Coefficients for the correlation between ratings of subjective confidence and test length.*

Examiner	Rho	p	Rho_{cc}	Df
A	.46	$< .001$	.59	48
B	.42	$< .01$	.50	34
C	.48	$< .001$	.54	48
D	.25	$< .05$	.30	48

The correlation coefficients are all in the direction hypothesized. There is a systematic tendency to positive correlation between ratings of subjective confidence and the length of the test stories. The coefficients are significant at the 5 per cent level at least for all the four examiners. A reasonable conclusion to be drawn is that the results strongly support our hypothesis that the test examiners in rating their subjective confidence are influenced by the length of the test stories.

C. Comments

It may be interesting to relate the findings to the test examiners' answers to the following question in a questionnaire after the

rating procedure: Do you consider that the number of words generally is a good indicator of the possibility of judging a story? The answers were: "To some degree. The longer a story is, the more modulated will the picture of the subject be. The story may however be too long, so that it is difficult to get a survey of it." (Examiner A)—"No." (Examiner B)—"Yes, in extreme cases. A very long or very short story may possibly, together with other test scorings, confirm a judgment." (Examiner C)—"In the present protocols there are some rather short, concise stories, which give more than some longer ones, but it seems to me that a positive relation exists between test length and the possibility of judging a protocol. Very short stories have been difficult to judge." (Examiner D.)

It will be observed that the examiner who most clearly states that there exists a positive relation, is examiner D, whose ratings of subjective confidence show the lowest correlation with test length.

The test examiners' more or less conscious reliance on the test length in rating their subjective confidence contributes to the systematic agreement between the test examiners on the individuals with respect to the subjective confidence with which they are judged (pp. 112–113).

2. OBJECTIVE ACCURACY AND TEST LENGTH

A. Problem

In the preceding section of the chapter we found a significant positive relation between the length of the test protocols and the subjective confidence with which they were judged. The next question is: How relevant is the test length as an indicator of the objective accuracy with which a subject is rated? As mentioned earlier (p. 56), Hartman found a tendency in a review of TAT records that the subjects most successfully judged were the least fluent in their TAT responses (Hartman, 1949, p. 30).

B. Procedure

We have found earlier that there are some individuals whose protocols are judged throughout over variables with more accuracy than others'. This holds good for all the examiners. There is

also substantial agreement between the test examiners about the subjects with respect to the accuracy with which they are judged (p. 115).

The relation between objective accuracy and test length has been estimated with rank order correlation for one group at a time for each examiner. The total accuracy score has been used as the estimate of each individual's position in the test examiner's distribution of accurately judged subjects. The rank order coefficients were transformed to one coefficient for each examiner by Fisher's z. The obtained coefficients are reported in Table 34.

TABLE 34. *Coefficients for the correlation between objective accuracy and test length*

Examiner	A	B	C	D	M
Rho	.08	−.11	.12	.16	.06

Df = 48

All the coefficients are insignificant. The mean coefficient is positive, but the examiner, whose ratings have the highest validity (B), obtains a negative coefficient. There is, however, no systematic relation between the coefficients in Table 34 and the mean validity coefficients for the test examiners.

Our result indicates no systematic relation between the possibility of judging a protocol validly and the length of the protocol.

C. Comments

The findings in this chapter indicate on the one hand that there is a significant positive correlation between the subjective confidence with which an individual is rated and the length of the test protocol. The correlation might be interpreted as an influence from test length on subjective confidence. On the other hand there is no indications of a systematic positive correlation between the possibility for a test examiner to judge a protocol validly and test length. The results imply that there is no actual basis for the test examiners' more or less conscious reliance on test length when judging the subjective confidence. The examiners are deceived by the length of the protocols. This fact may explain the lack of positive relationship between subjective confidence and objective accuracy over individuals (see p. 116).

Chapter 13. Test ratings and test length

A. Problem

The discussion about the role of the test length for the test examiners' confidence in their ratings led to another question. Does test length affect the test examiners' ratings of the students in the variables? Do long test stories lead to systematic ratings of any kind? If so, in which direction does the effect go?

B. Procedure and result

The distribution of words per test protocol is skewed. As the examiners' ratings were obtained with the group standard as norm, the distribution of number of words per protocol was transformed to a normal distribution with five categories for each group. The transformed scores obtained for test length were used for the computation of product-moment coefficients with the transformed ratings by the test examiners. The coefficients are reported in Table 35 for each of the experts. In the last column are given the mean coefficients for the four examiners. The variables are designated with the direction which we have called the "positive" direction, because of its positive correlation with sociometric status and on account of the criterion of reflection (pp. 48–49).

There is a systematic tendency to agreement between ratings in a certain direction and the length of the test stories. This may be illustrated by a further calculation for the coefficients in Table 35. The coefficients were rank ordered for each examiner and a coefficient of concordance computed for the agreement between the test examiners about the relationship between ratings in a certain direction and test length (Edwards, 1954, p. 402–14). The coefficient of concordance = .789, which gives a χ^2-value of 47.349. This value is well above significance at the 1 ‰ level. (Df = 15. χ^2 sign. at the 1‰ level = 37.697.) The reliability of the average rank order is expressed in a reliability coefficient of

TABLE 35. *Coefficients for the correlation between ratings made by the test examiners and test length*

Variable	A	p	B	p	C	p	D	p	$\bar{r}$
1. Intrapunitive	.06		.02		.11		.02		.05
2. Ostentatious	.42	<.001	.57	<.001	.41	<.01	.34	<.01	.44
3. Dominant	.41	<.01	.14		.38	<.01	.20		.28
4. Oppositional	.12		−.11		−.22		.00		−.05
6. Acute	.64	<.001	.62	<.001	.40	<.01	.39	<.01	.51
7. Well-adjusted	−.22		.18		.45	<.001	−.13		.07
9. Scrupulous	.19		.06		.06		.28	<.05	.15
10. Sensitive	.48	<.001	.53	<.001	.44	<.001	.43	<.001	.47
11. Independent	.35	<.01	.26	<.05	.26	<.05	.36	<.01	.31
12. Cheerful	.06		.21		.40	<.01	.05		.18
13. Objective	−.11		.19		.32	<.05	−.17		.06
14. Confident	.10		.37	<.01	.42	<.001	−.09		.20
15. Sociable	.50	<.001	.44	<.001	.43	<.001	.27	<.05	.41
17. Warmhearted	.39	<.01	.23		.36	<.01	.46	<.001	.36
18. Vital	.40	<.01	.68	<.001	.60	<.001	.55	<.001	.56
19. Tolerant	−.03		.05		.08		−.05		.01
M	.24		.27		.31		.18		

.911. The result means that the test examiners agree on the order of positive correlation with test length for different variables. In this meaning they agree to a significant degree.

We have found that ratings in a certain direction are systematically related to the length of the protocols, but that the objective accuracy with which a subject is rated is not systematically related to the length of the test protocol. If we use the definition by Rugg, quoted by Guilford (1954, p. 279) for the halo effect, "We judge our fellows in terms of a general mental attitude toward them; and there is, dominating this mental attitude toward the personality as a whole, a like mental attitude toward particular qualities", we may now call the tendency to systematic ratings in a certain direction a halo effect, with the length of the protocols as the basis.

Two possible interpretations of the results reported in Table 35 may be advanced. The first one is a tendency to ratings in the *positive* direction of a variable, when the ratings are based on long test protocols, and to ratings in the *negative* direction when

the ratings are based on short test protocols. This interpretation is supported by the mean coefficients in Table 35. All the mean coefficients, with one exception, are positive for the correlation between ratings in the direction of the variables which we have defined as the positive one and the test length. (For the exception, variable 4, Oppositional, there are other signs that indicate that the positive direction is the reverse for the test examiners to that for the students.)

The second interpretation is a tendency to *high* ratings for long test protocols, independent of the evaluation of the direction of the variables. The signs of the coefficients for the variables whose direction has been changed in Table 35 illustrate the discussion. In order to obtain the directions of the variables we have changed the direction for five variables, 1, 7, 13, 14 and 19. By this change of direction we changed the signs for twenty coefficients. Thirteen coefficients changed from negative to positive signs, seven coefficients changed from positive to negative signs. The mean coefficients, on which we based our first interpretation, are also all small. For these reasons we cannot exclude the interpretation that the coefficients in Table 35 may be the result of a tendency to high ratings, independent of the valuation of the high direction of the variables, and test length. Perhaps the two tendencies mentioned are combined. The coefficients for the variables whose directions have been changed favour this interpretation. The combination of the tendencies will lead to a strong positive relation between test length and ratings on those variables, where the "positive" and "high" directions coincide. For variables where the directions are not the same, the two tendencies will counteract each other and cancel out the effects.

C. Comments

Our finding that test length to a relatively high degree seems to affect both the subjective confidence with which a test examiner performs his ratings, and also the direction of the ratings, have practical applications of importance. In written examinations of the essay type the same factor of word number may play its role and affect the direction and the size of the ratings. It is also possible that a similar effect may influence ratings, based on other

sorts of information than written words. Does, for example, the amount of information that appears to have been obtained in an interview situation influence the ratings in any systematic way? A number of experimental studies on the role of irrelevant cues in ratings are reported from other situations, where the rating procedure has been applied (see e.g. Guilford, 1959, pp. 155–6). As Guilford points out, the effect of irrelevant cues is, of course, to reduce the validity of the ratings. Findings like those about the role of test length are of importance for rating procedures since knowledge of possible sources of error tends to diminish the effects (Guilford, 1954, p. 295).

Chapter 14.

Profiles for individuals rated with high or low confidence and for individuals rated with high and low objective accuracy

1. PROFILES FOR INDIVIDUALS WITH HIGH AND LOW CONFIDENCE RATINGS

On the basis of the results and interpretations accounted for in the preceding chapters, some further conclusions may be drawn and tested. This has been done for one aspect of the test ratings. Hitherto in the investigation relationships have been studied with correlation technique. In this section the relationships studied are given in the form of profiles in order to be more practically illustrative.

A. Problem

We have found that there exists a true rank order of individuals for each examiner with respect to the subjective confidence with which they are rated (p. 112). These rank orders were found to show positive correlation with test length, significant at the 5 per cent level at least for all the test examiners. We have also found a strong tendency to systematic relationships between test length and ratings in a certain direction of some variables (p. 123). From these results we can now predict that there ought to be a positive relation between ratings of great subjective confidence and ratings in the direction of the variables that correlate positively with test length. The prediction has been tested by studying the size and direction of the differences in ratings in various variables between two extreme groups of individuals, chosen so as to represent individuals, rated with high and low subjective confidence.

B. Procedure and results

With the combined ratings of subjective confidence from four test examiners for each individual as a basis, we obtained a distri-

bution of individuals. The reliability of this measure was estimated in a reliability coefficient of .618 (p. 113). Two extreme groups were now formed from the total distribution of confidence ratings. From 63 individuals in a normal distribution, eleven will fall above + 1 SD and eleven under — 1 SD from the mean. As the eleventh and twelfth student in both directions had the same obtained confidence scores, the extreme groups were composed of twelve individuals each. Thus we obtained groups for which the overlap of true scores is theoretically very small.

For each member of the groups the average rating was used as the subject's score on each variable. The mean of these average scores was calculated for each group on each variable. The computations were performed on the transformed scores. Thus the results are reported on a scale with the mean of 15.5 and the standard deviation of 5.0 for the total distribution.

The profiles based on the mean ratings obtained for each group on each variable are given in Fig. 7.

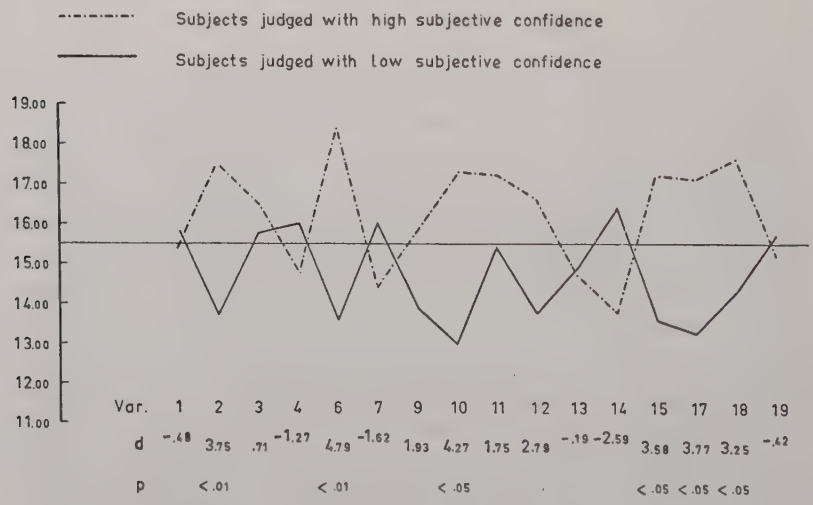


Fig. 7. Profiles for individuals with high and low confidence ratings.

Bartlett's test for the homogeneity of the variances (Guilford, 1956, pp. 242–4) showed that the variances for some variables were not homogeneous. Therefore the method proposed for such

circumstances by Behrens and given by Lindquist (1956, p. 97) was applied for testing the significance levels of the differences between the means of the extreme groups.

The differences are all in the direction that we predicted on the basis of the earlier results. Subjects judged with high subjective confidence are rated as significantly more ostentatious, acute, sensitive, sociable, warmhearted and vital than subjects judged with low subjective confidence. This means that for the variables for which test length correlated positively with high ratings by the examiners, the difference between the ratings of individuals with high confidence ratings, and the ratings of individuals with low confidence ratings, is positive. For those variables for which test length correlated positively with low examiners' ratings, the difference is negative.

C. Comments

The profiles support the first of our possible interpretations of the relationship in the preceding chapter, that the tendency is to a positive relation between long test stories and ratings in the "positive" direction of the variables.

It is sometimes argued that personality conflicts and disturbances are the easiest to judge from projective test protocols. If this holds good, for subjective confidence, we would have assumed a positive relation between high confidence ratings and ratings in the direction of the variables that indicate personality disturbances and inner conflicts. The results are contrary to such a tendency.

2. PROFILES FOR INDIVIDUALS JUDGED WITH HIGH OR LOW OBJECTIVE ACCURACY

A. Problem

In an earlier calculation we found that some individuals are rated with greater objective accuracy than others (pp. 115). A relevant and for the practical use of the test important question is: Does the profile for those individuals who are judged with great objective accuracy differ from the profile for individuals, judged with little accuracy?

B. Procedure and result

From the distribution of total accuracy scores for all the test examiners, two extreme groups of subjects were formed on the same principle as that for the extreme groups of subjects judged with various degrees of subjective confidence (p. 128). Thus each group will consist of eleven individuals.

The average of the ratings obtained from the four test examiners was calculated for each subject on each variable. The mean of these average scores was calculated for each group on each variable. The profiles, based on these values are given in Fig. 8.

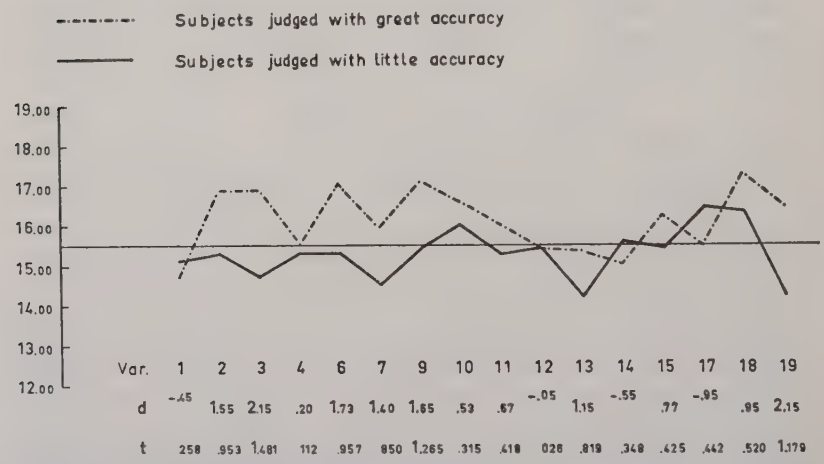


Fig. 8. Profiles for individuals judged with high and low objective accuracy.

For twelve variables the difference is positive. The mean ratings for subjects judged with great accuracy are higher than the mean ratings for subjects judged with little accuracy for these variables. None of the differences are, however, significant.

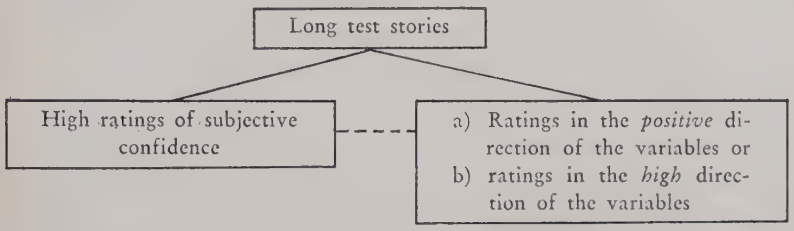
C. Comments

As mentioned earlier, it is sometimes argued that what is most easily judged from projective test protocols are indications of inner conflicts and personality disturbances. Our results do not indicate such a systematic relation between actual accuracy and the direction of the ratings.

3. TEST LENGTH, RATINGS OF SUBJECTIVE CONFIDENCE, AND TEST RATINGS

Here may be the place to summarize the findings in the last three chapters about the relationships between test length, ratings of subjective confidence, and test ratings.

Our results indicate a positive relation between long test stories, high ratings of subjective confidence and ratings in the "positive" (or high) direction of the variables.


Long test stories
High ratings of subjective confidence
a) Ratings in the *positive* direction of the variables or
b) ratings in the *high* direction of the variables

The basis of the relationships described above is most reasonably test length. The direction of the relationships may then be given different interpretations. It may be that the length of the test protocol for a certain subject affects directly both the ratings on the specific variables and the judgments of the subjective confidence with which the ratings are made. The two ratings would in this case be performed independent of each other. Another possible interpretation would be that long resp. short test protocols lead to ratings in a certain direction, which in turn leads to high resp. low ratings of subjective confidence, on account of psychological mechanisms within the test examiners. It is also possible that the test length affects the degree of subjective confidence, which in turn influences the ratings on the specific variables.

Chapter 15. Test length and students' ratings

A. Problem

We have found that for some variables the ratings made by the experts show significant correlation with test length. It may now be necessary to investigate whether there is any validity in these correlation coefficients in the meaning that the ratings by the students would also show correlation in the same direction with test length in the same variables.

B. Procedure and result

The coefficients for the correlation between the ratings made by the students and test length were computed as product-moment coefficients. For the students' ratings the scores on the transformed scale were used. The number of words per story was normalized on a five-point scale for each group (p. 123). The coefficients are reported in Table 36. The variables are designated with their "positive" directions for comparison with the coefficients in Table 35.

For the main aspects of variable 6, Acute—Dull, we have an objective measure in the result from the group test administered (p. 28). The test scores may be used as a criterion of the relevance of test length as an indicator of the individuals' positions in that variable. The product moment correlation between test scores and test length is .22.

C. Comments

Only two of the coefficients are significant at the 10 per cent level but we find the same tendency among these coefficients as among the coefficients for the relation between the test examiners' ratings and test length. Now we can conclude that part of the validity in the ratings made by the test examiners is due to the agreement between the students' ratings and test length. The con-

TABLE 36. *Coefficients for the correlation between the students' ratings and test length*

Variables	r	p
1. Intropunitive	–.02	
2. Ostentatious	.22	< .10
3. Dominant	.05	
4. Oppositional	–.09	
6. Acute	.08	
7. Well-adjusted	.11	
9. Scrupulous	.05	
10. Sensitive	.11	
11. Independent	.01	
12. Cheerful	.05	
13. Objective	.01	
14. Confident	.05	
15. Sociable	.05	
17. Warmhearted	.06	
18. Vital	.22	< .10
19. Tolerant	–.17	

N = 63 Df = 57

tribution to the validity is however very small, except for two of the variables; 2, Ostentatious and 18, Vital.

The tendency found is feasible. The same general "output" factor that generates long test stories from the individual may also to some degree lie behind and generate behavior. It is reasonable that above all such a factor may generate behavior, which is judged as vital and ostentatious by the fellow students.

Chapter 16.

Inter-rater reliability coefficients and validity coefficients with test length partialled out

A. Discussion

We have found that the test examiners with a high degree of agreement show significant coefficients for the correlation with test length in their ratings. We have also found that the students' ratings give correlation coefficients of the same direction in fourteen of sixteen variables. In this chapter we state our problem as follows: What is left of the inter-rater reliability coefficients if we partial out the variance that can be referred to test length? How much is left of the validity coefficients if we partial out the common variance that can be referred to test length?

The problem was studied by the calculation of partial correlation coefficients (Guilford, 1956, pp. 316–8).

B. Inter-rater reliability with test length partialled out

The result of the computations is shown in Table 37. For each test examiner in each variable the coefficients are the mean of the intercorrelation with the other examiners.

TABLE 37. *Inter-rater reliability coefficients with test length partialled out*

Variables	A	B	C	D	$\bar{r}$
1. Extrapunitive	.14	.24	.25	.23	.22
2. Ostentatious	.15	.27	.19	.10	.18
3. Dominant	.05	.14	.08	.12	.10
4. Oppositional	.28	.29	.29	.40	.32
6. Acute	.29	.24	.38	.30	.30
7. Maladjusted	.10	.22	.11	.22	.16
9. Scrupulous	.21	.19	.20	.12	.18
10. Sensitive	.34	.18	.23	.21	.24
11. Independent	.17	.14	.13	.18	.15
12. Cheerful	.25	.16	.25	.35	.25

13. Subjective	.10	.04	.00	.06	.05
14. Anxious	.18	.14	.21	.19	.18
15. Sociable	.10	.11	.18	.18	.14
17. Warmhearted	.43	.37	.49	.44	.43
18. Vital	.11	.16	.17	.22	.16
19. Intolerant	.26	.07	.24	.22	.20
Md	.175	.170	.205	.215	

N = 63

In the light of the previous discussion of inter-rater reliability, the coefficients in Table 37 may be considered as extremely low (pp. 81–84).

C. Validity coefficients with test length partialled out

In Table 38 are reported the validity coefficients with the test length partialled out.

TABLE 38. *Validity coefficients with test length partialled out*

Variable	A	p	B	p	C	p	D	p	$\bar{r}$
1. Extrapunitive	–.12		.05		.14		.01		.02
2. Ostentatious	.03		.09		.02		–.01		.03
3. Dominant	–.02		.01		.17		.12		.07
4. Oppositional	.15		–.01		.17		.11		.11
6. Acute	–.12		.36	<.01	.18		.23	<.10	.16
7. Maladjusted	–.14		.21		.15		.14		.03
9. Scrupulous	.09		.20		–.15		–.03		.03
10. Sensitive	.19		.33	<.01	.05		.50	<.001	.27
11. Independent	–.04		.33	<.01	.11		.21		.15
12. Cheerful	.10		.34	<.01	.14		.39	<.01	.24
13. Subjective	–.24	<.10	–.10		–.12		–.04		–.13
14. Anxious	–.26	<.05	.24	<.10	.15		.28	<.05	.10
15. Sociable	.16		.11		.34	<.01	.29	<.05	.23
17. Warmhearted	.24	<.10	.33	<.01	.17		.28	<.05	.26
18. Vital	–.27	<.05	.02		–.22	<.10	.00		–.12
19. Intolerant	–.04		.38	<.01	.06		.23		.16

The validity coefficients corrected for attenuation in the criterion have been used for the calculation of the coefficients.

As we can see in a comparison with the values in Table 23 (p.

90), the lowering effect is not great. This is due to the fact that for most variables there is very low correlation between the students' ratings and test length. And for the two variables, for which there are correlations of any importance between the students' ratings and test length, there are negligible coefficients for the correlation between the examiners' ratings and test length.

Besides the students' ratings of variable 6, we have also used another criterion for the test ratings in this variable. With the group test administered as a criterion, we obtained validity coefficients of .18, .48, .31 and .17 respectively for A, B, C and D. We have also found that the coefficient for the correlation between test length in number of words for the TAT and test scores on the group test is .22. If we partial out the test length in the validity coefficients reported we obtain the following: .05 (A), .44 (B), .25 (C) and .09 (D).

Chapter 17.
TAT versus Rorschach and Sentence completion

In the investigation of which this study is a part, the Rorschach test and a Sentence completion test were also included (p. 18). (As examples of investigations in this area see e.g. More, 1957, Shatin, 1953, pp. 92–101 or Shatin, 1955, pp. 317–31. For a general discussion of the relationship between various projective tests see e.g. Shneidman, 1956, pp. 595–642). For the sixteen variables that have been rated by the TAT examiners, we also have ratings based on the Rorschach, obtained from Rorschach experts who have worked with the same instruction as the TAT examiners. The items in the Sentence completion test were constructed to give materials for ratings in nine variables.

1. RORSCHACH

A. Ratings, based on the Rorschach protocols

The ratings of the Rorschach protocols have been performed by two test examiners for each group. Each examiner has rated two groups but never in combination with the same expert for both the groups. Thus every protocol has been rated by two examiners. When we want to study the relation of our ratings based on the TAT to ratings based on Rorschach protocols, the most reliable estimate of an individual's position in ratings from the Rorschach is the pooled ratings from the two test examiners, who have rated him.

The ratings for each examiner and each group were transformed to standard scores. Then the standard scores for each subject were summarized. The summarized scores were transformed to a distribution with the same mean (15.5) and the same standard deviation (5.0) as the distribution of the transformed TAT ratings.

The reliability of the ratings based on the Rorschach was com-

puted as inter-rater reliability. The mean inter-rater reliability coefficients for each variable have been used in the computation of the coefficients for the reliability in the pooled ratings by means of Spearman–Brown's formula. The inter-rater reliability coefficients for the pooled ratings are reported in Table 39.

TABLE 39. *Reliability of pooled ratings based on the Rorschach*

Var.	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
r_{tt}	.63	.72	.73	.53	.68	.45	.66	.63	.57	.56	.60	.64	.67	.44	.72	.54

B. TAT—Rorschach

The relation of the TAT ratings made by one examiner at a time to the pooled ratings on the Rorschach was estimated by product-moment coefficients over all individuals on the transformed scores. The coefficients corrected for attenuation in the Rorschach ratings by means of the reliability coefficients in Table 39 are reported in Table 40.

TABLE 40. *Coefficients for the correlation between TAT-ratings and pooled ratings on the Rorschach (coefficients corrected for attenuation)*

Variable	A	p	B	p	C	p	D	p	$\bar{r}$
1. Extrapunitive	.00		.21		−.06		.14		.07
2. Ostentatious	.10		.19		.14		.06		.12
3. Dominant	.20		.31	<.05	.18		.10		.20
4. Oppositional	.42	<.001	.35	<.01	.27	<.05	.18		.31
6. Acute	−.25	<.10	.05		.13		−.23		−.07
7. Maladjusted	.38	<.01	.02		.07		.05		.13
9. Scrupulous	.24	<.10	.18		.05		.31	<.05	.20
19. Sensitive	−.06		.04		−.05		−.06		−.03
11. Independent	−.04		.24	<.10	.28	<.05	.16		.16
12. Cheerful	−.11		.10		−.24	<.10	−.37	<.01	−.16
13. Subjective	.24	<.10	.01		.17		.16		.14
14. Anxious	.09		.08		.21		−.15		.06
15. Sociable	−.07		−.20		−.34	<.01	−.08		−.17
17. Warmhearted	−.25	<.10	−.30	<.05	−.29	<.05	−.42	<.001	−.32
18. Vital	.00		−.01		−.04		−.33	<.01	−.09
19. Intolerant	−.12		.10		.02		−.21		−.05
Md	.00		.09		.06		.00		.065

N = 63 Df = 57

C. Comments

The mean coefficients give one positive and one negative coeffient significant at the 5 per cent level. For one of these variables, Oppositional, three test examiners show significant coefficients at the 5 per cent level at least. On the other hand all the test examiners obtain negative coefficients which are significant at the 10 per cent level at least for variable 17, Warmhearted. The result indicates no systematic positive relations of the ratings based on the TAT to the pooled ratings based on the Rorschach when both types of ratings have been obtained as predictions of ratings of observed behavior. This may only mean that the TAT ratings do not cover the same area of information as the Rorschach. The validity coefficients calculated for the ratings based on the Rorschach are of the same level as those for the TAT ratings. Thus the situation is favourable for the complementary use of the two tests.

2. SENTENCE COMPLETION

A. Variables

As no Swedish version of the Sentence completion test had earlier been standardized, a new form of the test was constructed by Derestad (1958) for use in the investigation. The test, in the following called the SCD, consists of 71 items. The items were constructed to give material for ratings in nine of the variables that were common to the tests in the investigation. The following variables were intended to be rated from the test material:

1. Intropunitive — Extrapunitive
2. Reticent — Ostentatious
3. Submissive — Dominant
4. Believing in authority — Oppositional
7. Well-adjusted — Maladjusted
9. Unscrupulous — Scrupulous
12. Gloomy — Cheerful
14. Confident — Anxious
15. Unsociable — Sociable

B. Ratings, besed on the SCD

The test protocols have been obtained from the same students as for the TAT. Sentence completion data are however missing for five students. Thus the number of Sentence completion protocols for comparison with the TAT data is 58.

Two examiners have rated the test protocols with two different instructions. First they rated independently of each other one item at a time in the variable, for which the item was intended. The inter-rater reliability was very high for these ratings. The coefficients for the reliability of the items in each variable, obtained by means of Spearman–Brown's formula from the mean inter-rater reliability coefficient for each item, are reported in Table 41. For each variable the ratings were performed on six items.

TABLE 41. *Inter-rater reliability of pooled ratings, based on the SCD (Ratings of items)*

Variable	1	2	3	4	9	12	14	15
r_{tt}	.967	.983	.969	.978	.978	.985	.978	.985

The pooled ratings in these variables, obtained by summing up the standard scores, were used as estimates of the individuals' positions according to the SCD.

No ratings have been obtained with this instruction for variable 7, Maladjusted, as no special items were constructed for this variable.

When the test examiners had rated all the protocols according to the first instruction, a new rating procedure was performed with another instruction. Now the test examiners, independently of each other, made an over-all rating of the test protocol for each individual in each variable. The whole test was thus used as a basis for the ratings with this instruction. In this case variable 7, Maladjusted, was also rated.

The pooled ratings from the two test examiners, obtained by summing up the standard scores, were used as the best estimate of an individual's position in the distribution of ratings based on the SCD. The reliability of the pooled ratings, computed by means of

Spearman–Brown's formula from the inter-rater reliability coefficients, are reported in Table 42.

TABLE 42. *Inter-rater reliability of pooled ratings based on the SCD (over-all rating)*

Variables	1	2	3	4	7	9	12	14	15
r_{tt}	.56	.77	.51	.35	.73	.71	.79	.67	.64

C. TAT—SCD

The relation of the TAT ratings for each test examiner to the ratings based on the SCD was estimated by product-moment correlation. The coefficients are reported in Tables 43 and 44. The coefficients in the last table are corrected for attenuation in the ratings from the Sentence completion test (Table 41).

TABLE 43. *Coefficients for the correlation between ratings, based on the TAT and on the SCD (ratings of items)*

Variable	A	B	C	D	$\bar{r}$
1. Extrapunitive	.12	−.08	−.03	−.03	.00
2. Ostentatious	.02	.10	.22	.01	.09
3. Dominant	−.07	.05	−.09	−.08	−.05
4. Oppositional	−.18	.34	.09	.03	.07
9. Scrupulous	.11	−.01	.20	.20	.13
12. Cheerful	−.04	.20	.04	.02	.05
14. Anxious	.28	.21	.16	.14	.19
15. Sociable	.07	.05	.19	.00	.08

N = 58 Df = 52

Even if only one of the mean coefficients is negative, the tendency to agreement between the TAT ratings and ratings based on the SCD, when these last are obtained in ratings of one item at a time, is very low. Only two coefficients (one for examiner A and one for examiner B) are significant at the 5 per cent level. This is what may be expected by chance in a table with thirty-two coefficients.

TABLE 44. *Coefficients for the correlation between the TAT ratings and pooled ratings from the SCD (over-all ratings)*

Variable	A	p	B	p	C	p	D	p	$\bar{r}$
1. Extrapunitive	−.01		−.02		−.12		−.21		−.09
2. Ostentatious	.00		.09		.31	<.05	−.01		.10
3. Dominant	.09		.19		.00		−.09		.05
4. Oppositional	−.06		.27	<.10	.07		.24		.13
7. Maladjusted	.31	<.05	.38	<.01	.35	<.01	.27	<.10	.33
9. Scrupulous	.09		−.12		.43	<.01	.17		.15
12. Cheerful	.14		.21		.16		.29	<.05	.20
14. Anxious	.29	<.05	.13		.19		.18		.20
15. Sociable	.12		.42	<.01	.12		.16		.21

N = 58 Df = 52

Even if the coefficients are not all positive there is a clear tendency for the TAT ratings to agree with ratings based on the SCD, when the SCD ratings are performed as over-all ratings. The agreement is marked for variable 7, Maladjusted, for which all the test examiners have obtained significant coefficients for the relation to the SCD. If we consider the whole set of coefficients, we find that for the variables 7, Maladjusted, 12, Cheerful, 14, Anxious, and 15, Sociable, all the coefficients are positive and at least one test examiner per variable has obtained a significant coefficient. Those four variables have significant intercorrelation coefficients, with two exceptions, in the correlation matrices for all the four examiners. The exception is that the coefficients for the correlation between variables 7 and 15 are not significant for examiners A and D. The negative directions of variables 7, 12, and 14, (Maladjusted, Gloomy, Anxious), indicate personality disturbances. They also correspond most closely to the content of the "alienation syndrome" earlier shown to be predictable with success from the TAT.

PART IV

Summary and suggestions for further research

Chapter 18. Summary

The study reported here is concerned with the reliability and validity of ratings, based on TAT protocols and obtained by blind analysis. The reliability studied is inter-rater reliability. The concurrent validity of the method is studied by the correlation between ratings of peers made by fellow students on a number of personality variables and the predictions of these ratings from the TAT protocols. By relating the predictions of the students' ratings from the TAT protocols to the predictions based on two other projective methods the construct validity of the method is studied. The studying of some factors influencing the ratings may also be expected as relevant for the question of construct validity.

The study is part of an investigation with four other projective methods, for which separate accounts are delivered. This fact has involved some limitations in the plans for the TAT study.

Part I, chapters 1–3, is concerned with the criterion data. The subjects in the TAT investigation were five groups of male students with a mean age of about twenty-one years from four Swedish People's High Schools. All the students in each group have rated all other students in the group for twenty variables (definitions given in Appendix B). Precautions were taken in order to minimize possible stereotype effects. The reliability of the ratings made by the students was estimated as inter-rater reliability by split-half technique. The median of the reliability coefficients found in this way was .855. A group intelligence test was administered in order to gauge the intellectual level of the students.

For the study of the structure in the correlation matrix for the students' ratings a factor analysis and a cluster analysis have been performed. Both show that there exists a structure with factors which seem meaningful. The hypothesis that one of the factors is due to halo to some degree is studied by a method reminiscent of Eysenck's criterion analysis. The results support the hypothesis.

Part II, chapters 4–6, accounts for the projective test data. TAT stories of fifteen TAT pictures, obtained from 63 subjects of the 65, who were the total number of students in the groups investigated, were interpreted by blind analysis on sixteen of the variables on which the students had rated each other. The test examiners were four clinical psychologists with experience of the method in clinical work. No special scoring system was prescribed for use. Each test examiner followed the method he had found the most useful. Besides the ratings, performed as predictions of the students' ratings, the test examiners were also given the opportunity to rate the latent tendencies of a subject, which may be suppressed and not expressed in observable behavior. This possibility was used by one examiner consistently. For each rating delivered, the test examiners also judged the subjective confidence with which the ratings were performed.

In order to study the structure of the test examiners' ratings two methods were applied, a cluster analysis and the correlation between the coefficients in one column of the matrix for each examiner's ratings at a time with the same column of coefficients for each of the other test examiners. The agreement in structure between the test examiners mutually was found to be higher than that between the students' ratings and the ratings by each of the test examiners.

In part III the main results are given.

Chapter 7 begins with a general discussion of the error sources in inter-rater reliability coefficients. Attention is drawn to the fact that the lack of agreement between the examiners is the only source of error determining the deviation from a reliability coefficient of 1.0. From this point of view the inter-rater reliability coefficients may be expected as maximum values for the reliability. The inter-rater reliability coefficients for the ratings in the present investigation are reported. The mean inter-rater reliability is distressingly low, .29. Some circumstances necessary to remember for the adequate valuation of the result are mentioned.

Chapter 8 accounts for the results of the study of concurrent validity. With consideration taken to the circumstances to which attention is drawn, the results are interpreted as promising. Except for one variable at a time the validity was also studied for the

"alienation syndrome" which was found to be predictable with success from the TAT. No systematic relation between the size of the validity coefficients and the degree of agreement in structure between the students and each of the test examiners was found.

In chapter 9 some suggestions for the practical use of the test are discussed on the basis of the results found for the inter-rater reliability and the validity of the ratings based on the TAT.

In chapter 10 is investigated the question of whether there exists a true rank order between the test examiners with respect to the ability to predict the students' ratings and if a true rank order exists between the variables, in the meaning that some variables are judged with greater accuracy than others throughout. The results support an affirmative answer to these questions.

Chapter 11 discusses the subjective confidence with which each prediction has been made and its relation to the actual accuracy of the rating. Attention is drawn to the fact that the general problem of subjective confidence has two aspects; one dealing with the subjective confidence over variables for one individual at a time, the other with the subjective confidence over individuals for one variable at a time. It is found that there exists a true rank order between the subjects with respect to the subjective confidence with which they are rated. This holds good for each of the test examiners. The examiners also agree about the individuals in this respect. A measure of the accuracy with which each subject is rated is applied. It shows that there are some individuals, who are predicted with greater accuracy than others throughout. Also in this respect the examiners agree. When the two measures of subjective confidence and objective accuracy were related to each other, we found no systematic correlation. Nor was there found any systematic correlation between subjective confidence and objective accuracy over variables.

Starting with the hypothesis that the amount of information experienced is the main basis of subjective confidence, we related the measure of subjective confidence to the test length in chapter 12. Then we obtained significant positive coefficients for the correlation between the subjective confidence with which the ratings are delivered and the length of the test protocols. When we related the measure of objective accuracy to test length, we found

no significant coefficient for the correlation and no systematic tendency.

In chapter 13 is investigated the role of test length for the ratings based on the TAT. There is found a systematic positive correlation between test length and ratings in a certain direction of the variables. Two possible interpretations are developed; a) a tendency to ratings in the *positively evaluated* direction of the variables or b) to ratings in the *high* direction of the scale when the test stories on which the ratings are based are long.

On the basis of the results in chapter 11–13 some conclusions are drawn and tested in chapter 14. The calculations are performed for two extreme groups and the results are given in profiles. It is found that subjects judged with high subjective confidence are rated higher in the "positive" direction of the variables than individuals judged with less subjective confidence. The differences between the two groups are significant at the 5 per cent level at least for six variables, when two tail tests of the significance level are applied. Subjects judged with high subjective confidence are rated as significantly more ostentatious, acute, sensitive, sociable, warmhearted and vital than subjects judged with less subjective confidence. The profiles calculated for two extreme groups rated with high and low objective accuracy show no significant differences.

In chapter 15 the relation between the students' ratings in various variables and test length is investigated. The coefficients for the correlation have the same signs as those for the correlation between the test examiners' ratings and test length for fourteen of the sixteen variables. The coefficients are however very low for all the variables except two, Ostentatious and Vital.

In order to estimate the inter-rater reliability coefficients and the validity coefficients, when the role of test length is partialled out, partial correlation coefficients were calculated and reported in chapter 16. On account of the fact that the coefficients for the correlation between the students' ratings and test length were very small for most variables, the lowering effect on the validity coefficients of partialling out the role of test length was small. The inter-rater reliability coefficients were, however, lowered to median coefficients of about .20.

The relation to ratings based on the Rorschach and the Sentence completion test is reported in chapter 17. The median coefficient for the correlation between the ratings based on the TAT and the Rorschach was .065. Seven of the coefficients are negative and nine positive.

The ratings based on a Swedish form of the Sentence completion test (SCD) were obtained by two instructions; a) by rating one item at a time and b) by using the whole test for the rating of each subject in the specific variables. For the first type of SCD ratings no systematic positive correlation was found with the TAT ratings. For the correlation between the ratings based on the TAT and the ratings obtained by using the whole Sentence completion protocol as a basis for the ratings, mean positive coefficients were obtained for eight of the nine variables studied. For the variables which most closely correspond to those forming the alienation syndrome, defined by Davids, all the test examiners show positive coefficients for the correlation with over-all ratings of SCD. For the main variable of the syndrome (Maladjusted) all the examiners obtain coefficients significant at the 10 per cent level at least.

Chapter 19. Suggestions for further research

On the basis of the results obtained in this study, some suggestions for further research on the reliability and validity of the TAT technique may be made.

The situation is, as pointed out in an earlier connection (p. 102), that we have test examiners, each of whom with somewhat different methods, makes use of some relevant information but not necessarily the same as the others. In this situation it is of dubious value to increase the inter-rater reliability only by giving detailed instructions for the judging, as is done in many investigations. It is true that such instructions will have the effect of decreasing the subjective component of the ratings and of raising the inter-rater reliability. There are also examples of very high coefficients for inter-rater reliability obtained in such a way. The profit is, however, imaginary. The only advantage of a detailed instruction, e.g. by choosing one of the methods used by the test examiners in this investigation, would be that the test examiners would use the same information in the observed material and rate it in the same way. Such a procedure will affect the validity of the ratings only under certain conditions. High inter-rater reliability is not of practical importance, unless it is followed by validity.

Our finding that various test examiners can find some but not necessarily the same relevant information as a basis for valid ratings indicates tasks for further research. One task should be to find out the kind of information that various methods and various theories use in the test protocol for the interpretation and which information the different methods and theories can use for valid interpretations. This task is very difficult. If it were possible and we could use all our knowledge to form a new instruction covering all relevant information we should, however, increase considerably the usefulness of the TAT for assessing personality variables.

Our results indicate that there exists a true rank order between the test examiners in their ability to predict the students' ratings. This suggests another task for future research, namely to study which features are typical of a good test examiner and to estimate the significance of the training of examiners in the practical use of the technique. Murray stated 1943 (p. 6): "The future of the TAT hangs on the possibility of perfecting the interpreter (psychology's forgotten instrument) more than it does on perfecting the material."

The validity coefficients reported in Table 23 (p. 90) for the predictions of the students' ratings, are calculated on ratings based on samples of behavior from one occassion only. This fact introduces the problems of function fluctuation and time sampling. These problems are of importance not only for the discussion of reliability (see Ekman, 1947) but also for the question of validity (see Björkman, 1953). Data presented by Björkman (1953, p. 34) indicate that the problems are of less importance for the validity of an ability test. It will, however, be of greater importance in questions of projective techniques. This for two reasons: a) Factors bound to the specific test situation may be expected to be of greater influence and b) systematic, periodical and chance fluctuations within the individual (see Ekman, 1947, p. 255) may be expected to be greater.

Two possible solutions of the problems of function fluctuation and time sampling may be mentioned. First, the total sampling of behavior from more than one occasion could be used as a basis for one rating of each subject. Thus it should be possible for the test examiner to take into consideration divergencies between the contents of the test protocols from the different occasions for each subject. Second, the sampling from each of several successive test sessions should be used as a basis for the rating of the subject. The pooled ratings from the different occasions could then be used as the best estimation of the subject's positions on the variables judged. If a new sample of behavior from another occasion should have the same validity as the first, the validity of the combined measures will depend on the degree of intercorrelation between the two series of predictions. The size of the intercorrela-

tions is in turn dependent on the reliability of the ratings on each occasion and the extent of the effects of the factors mentioned above.

The effect of both these procedures would be to increase the validity of the ratings based on the total amount of information available. The extent of the effect and its relation to the time and money expended may be investigated in order to determine how the law of diminishing returns functions in this case.

Appendices and references

APPENDIX A

TABLE 45. *Correlation matrix for the students' ratings I (20 × 20 variables. N = 83)*

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	.641	.602	.480	.554	−.252	.366	−.339	−.615	−.023	−.041	−.034	.681	.243	.288	−.383	−.604	−.387	.790	.224
2		.801	.518	.690	.033	.137	−.183	−.481	.218	.217	.349	.370	−.147	.649	−.219	−.266	−.082	.430	.326
3			.746	.429	.313	.224	−.055	−.320	.332	.533	.191	.208	−.246	.525	.017	−.213	.105	.313	.276
4				.399	.249	.432	−.151	−.421	.299	.519	.016	.149	−.105	.267	.074	−.315	.060	.240	.273
5					−.416	.077	−.498	−.785	−.113	−.283	.358	.575	.259	.606	−.625	−.397	−.413	.559	.320
6						−.019	.390	.438	.553	.833	.013	−.622	−.691	.026	.739	.282	.709	−.504	−.126
7							−.197	−.172	.249	.245	−.537	.409	.323	−.253	.287	−.163	−.233	.348	.581
8								.643	.387	.319	.115	−.280	−.260	−.031	.334	.389	.502	−.320	−.089
9									.244	.340	−.204	−.490	−.251	−.350	.592	.549	.463	−.472	−.163
10										.579	.036	−.202	−.264	.113	.518	.235	.440	−.207	.375
11											−.113	−.407	−.523	.019	.709	.133	.580	−.315	.110
12												−.060	−.233	.804	−.401	.195	.010	−.119	−.023
13													.658	.091	−.494	−.425	−.532	.760	.362
14														−.208	−.420	−.258	−.513	.518	.284
15															−.375	.134	.067	.113	.195
16																.359	.603	−.478	.031
17																	.406	−.505	.132
18																		−.494	−.172
19																			.184

TABLE 46. *Correlation matrix for the students' ratings II (16 × 16 variables. N = 65)*

	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
1	.496	.535	.347	−.141	.245	−.486	−.065	−.087	−.064	.592	303	.239	−.419	−.255	.702
2		.791	.379	.128	−.030	−.419	.128	.167	.402	.299	−.249	.658	−.062	.027	.311
3			.610	.371	.025	−.312	.191	.472	.255	.126	−.256	.588	−.022	.105	.260
4				.327	.231	−.264	.226	.478	.037	−.091	−.147	.291	−.132	.146	.200
6					−.034	.435	.481	.859	−.035	−.582	−.585	.067	.217	.652	−.293
7						−.126	.091	.068	−.616	.282	.321	−.391	−.050	−.261	.286
9							.307	.374	−.200	−.413	−.218	−.367	.299	.502	−.361
10								.416	.009	−.285	−.169	.133	.133	.422	−.076
11									−.050	−.501	−.503	.021	.201	.566	−.226
12										.013	−.251	.756	.083	.195	−.164
13											.547	.127	−.341	−.429	.632
14												−.247	−.225	−.324	.453
15													.118	.097	.036
17														.153	−.475
18															−.305

TABLE 47. *Centroid matrix*

Variable	I	II	III	IV	V	VI	h^2
1	.771	−.195	−.319	−.217	−.240	−.032	.840
2	.596	−.651	.069	−.098	−.177	.154	.849
3	.388	−.763	−.153	−.241	−.226	.032	.866
4	.360	−.647	−.291	−.239	.136	−.150	.734
5	.856	−.228	.274	−.036	.114	.142	.894
6	−.632	−.650	−.112	−.239	−.044	−.074	.899
7	.255	−.152	−.739	.141	.340	.178	.802
8	−.550	−.166	−.067	.292	−.409	−.122	.602
9	−.819	.072	−.174	.237	−.289	.138	.865
10	−.253	−.638	−.284	.259	.046	−.074	.626
11	−.406	−.735	−.374	−.212	−.025	−.026	.892
12	.173	−.323	.706	.248	−.256	−.241	.818
13	.786	.171	−.317	.196	−.155	.162	.835
14	.485	.503	−.310	.328	.060	.036	.697
15	.401	−.585	.588	.188	−.256	.083	.957
16	−.708	−.360	−.443	−.078	.172	.197	.902
17	−.528	−.142	.213	.496	−.030	.217	.638
18	−.651	−.447	.054	−.100	−.117	.118	.665
19	.767	.130	−.319	−.078	−.278	.142	.810
20	.325	−.336	−.286	.558	.302	.197	.742

TABLE 48. *Rotated matrix (Orthogonal rotation)*

Variable	A	B	C	D	E	F	h^2
1	.519	−.557	−.048	.307	−.402	−.032	.839
2	.172	−.682	.424	.342	−.186	.154	.850
3	−.084	−.774	.225	.340	−.307	.032	.868
4	−.061	−.692	−.035	.472	−.038	−.158	.733
5	.624	−.480	.385	.268	.152	.142	.894
6	−.886	−.326	−.014	−.012	−.049	−.074	.899
7	.063	−.086	−.478	.725	−.054	.178	.800
8	−.559	.310	.175	−.051	−.380	−.122	.601
9	−.674	.514	−.104	−.184	−.288	.138	.865
10	−.575	−.107	.106	.506	−.104	−.074	.626
11	−.765	−.442	−.127	.258	−.168	−.036	.893
12	.048	−.083	.859	−.107	.041	−.241	.818
13	.724	−.050	−.040	.406	−.342	.162	.836
14	.642	.355	−.227	.299	−.123	.036	.696
15	.093	−.368	.891	.114	−.032	.083	.957
16	−.832	−.010	−.381	.164	.003	.197	.902
17	−.500	.461	.348	.044	.067	.217	.638
18	−.779	−.099	.104	−.141	−.042	.118	.664
19	.686	−.261	−.109	.225	−.434	.142	.810
20	.072	.072	.130	.819	.072	.197	.742

TABLE 49. *Transformation matrix (Orthogonal rotation)*

	A	B	C	D	E	F
I	.8473	−.3706	.1660	.3357	−.0724	0000
II	.5215	.6022	−.4046	−.4478	.0339	0000
III	.0995	0000	.7316	−.5119	.4390	0000
IV	0000	.7071	.4300	.5573	−.0668	0000
V	0000	0000	−.2991	.3378	.8924	0000
VI	0000	0000	0000	0000	0000	1.0000

TABLE 50. *Transformation matrix (Oblique rotation)*

	A_1	B_1	C_1	D_1	E_1	F_1
A	−.611	−.056	.056	−.040	−.227	−.025
B	−.644	−.652	−.097	.390	.427	−.110
C	−.132	−.213	.975	.091	.194	.056
D	.037	−.136	−.124	.827	.042	−.204
E	.414	−.431	.047	.029	−.776	−.042
F	−.146	−.568	−.136	−.392	−.359	.970

TABLE 51. *Cosines between reference vectors*

	A_1	B_1	C_1	D_1	E_1	F_1
A_1	1.00	.38	−.07	−.14	−.43	−.09
B_1		1.00	−.07	−.17	.23	−.44
C_1			1.00	.00	.14	−.04
D_1				1.00	.35	−.59
E_1					1.00	−.35
F_1						1.00

TABLE 52. *Correlation matrix. Examiner A (N = 63)*

	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
1		.146	.311	.686	−.215	.133	−.200	−.405	−.151	−.021	.232	.393	−.008	−.453	.076	.685
2			.729	.318	.380	.084	−.096	.253	.523	.415	−.081	−.309	.562	.200	.711	.105
3				.568	.249	.103	−.174	.079	.538	.272	−.029	−.332	.538	.023	.639	.244
4					−.008	.166	−.289	−.259	.160	−.002	.187	.078	.241	−.332	.268	.468
6						.327	.382	.751	.417	.050	.088	−.227	.297	.628	.253	−.238
7							.542	.272	−.181	−.590	.719	.399	−.078	.111	−.195	.345
9								.532	−.207	−.313	.468	.070	−.170	.491	−.194	−.015
10									.254	−.031	.091	−.232	.149	.831	.098	−.432
11										.432	−.361	−.614	.537	.226	.709	−.271
12											−.503	−.534	.512	.063	.551	−.243
13												.418	−.112	.009	−.274	.437
14													−.409	−.311	−.478	.547
15														.208	.623	−.053
17															.136	−.481
18																.006
19																

TABLE 53. *Correlation matrix. Examiner B (N = 63)*

	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
1		.210	.504	.501	.094	−.092	.074	−.078	.115	.017	.032	−.175	.127	−.370	.089	.183
2			.462	.101	.622	−.248	.062	.410	.510	.299	−.301	−.506	.703	−.034	.694	−.091
3				.319	.369	−.108	−.070	−.072	.391	.126	−.226	−.416	.318	−.317	.400	.172
4					.022	.113	.116	−.195	.122	.029	.078	−.086	−.067	−.513	−.044	.199
6						−.404	−.109	.368	.589	.417	−.380	−.654	.656	.095	.672	−.176
7							.650	−.283	−.525	−.702	.754	.685	−.452	−.148	−.276	.542
9								−.231	−.257	−.520	.554	.391	−.189	−.254	−.090	.647
10									.244	.248	−.151	−.245	.439	.345	.290	−.344
11										.365	−.543	−.707	.434	.064	.549	−.197
12											−.614	−.527	.515	.092	.384	−.536
13												.606	−.399	−.248	−.363	.535
14													−.632	.013	−.578	.282
15														.228	.716	−.220
17															.259	−.230
18																−.045
19																

TABLE 54. *Correlation matrix. Examiner C (N = 63)*

	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
1		.199	.089	.503	−.120	.379	−.413	−.563	−.060	−.256	.283	.212	−.254	−.594	−.227	.456
2			.728	−.134	.516	−.342	−.157	.138	.441	.320	−.346	−.483	.462	.088	.441	−.005
3				−.030	.679	−.320	−.140	.142	.629	.310	−.376	−.573	.339	.082	.457	−.116
4					−.171	.618	−.107	−.478	−.071	−.509	.444	.263	−.461	−.606	−.445	.487
6						−.542	−.079	.466	.665	.396	−.621	−.821	.467	.352	.618	−.350
7							−.044	−.519	−.265	−.815	.743	.596	−.765	−.646	−.742	.518
9								.243	−.085	−.072	−.021	.043	.129	.264	.150	−.025
10									.310	.503	−.523	−.416	.572	.744	.600	−.620
11										.215	−.453	−.538	.252	.193	.348	−.311
12											−.599	−.477	.790	.513	.669	−.364
13												.630	−.633	−.500	−.675	.609
14													−.472	−.340	−.616	.330
15														.590	.683	−.441
17															.600	−.727
18																−.420
19																

TABLE 55. *Correlation matrix. Examiner D (N = 63)*

	1	2	3	4	6	7	9	10	11	12	13	14	15	17	18	19
1		.190	.166	.409	−.009	−.031	−.072	−.130	.202	−.113	−.065	.053	−.213	−.198	.024	.332
2			.389	−.013	.345	−.147	.011	.131	.395	.287	.062	−.358	.506	.367	.539	−.101
3				.185	.269	.106	.117	.162	.398	−.014	−.059	−.222	.222	.136	.330	.063
4					.029	.338	−.218	−.046	.127	−.342	−.048	.192	−.255	−.381	−.083	.221
6						−.075	.226	.581	.623	.086	−.046	−.341	.245	.362	.531	−.061
7							.143	.092	.139	−.651	.286	.579	−.236	−.186	−.189	.166
9								.370	.166	−.082	.286	.047	−.048	.395	.373	−.010
10									.417	−.053	.241	−.030	.251	.321	.392	.081
11										−.042	.067	−.356	.156	.281	.458	.051
12											−.226	−.589	.406	.483	.289	−.314
13												.338	−.009	.051	−.087	.194
14													−.380	−.371	−.392	.228
15														.457	.378	−.254
17															.628	−.347
18																−.043
19																

TABLE 56. *Clusters for examiner A*

Cluster			Variable			M_r
I	1	(10)	Insensitive	—	Sensitive	.70
	2	(17)	Cold	—	Warmhearted	.65
	3	(6)	Dull	—	Acute	.52
	4	(9)	Unscrupulous	—	Scrupulous	.47
II	1	(18)	Lacking vitality	—	Vital	.62
	2	(11)	Dependent	—	Independent	.56
	3	(2)	Reticent	—	Ostentatious	.54
	4	(15)	Unsociable	—	Sociable	.53
	5	(3)	Submissive	—	Dominant	.51
	6	(12)	Gloomy	—	Cheerful	.45
	7	(14)	Anxious	—	Confident	.45
III	1	(1)	Extrapunitive	—	Intropunitive	.56
	2	(17)	Cold	—	Warmhearted	.52
	3	(19)	Intolerant	—	Tolerant	.52
	4	(10)	Insensitive	—	Sensitive	.48
	5	(4)	Oppositional	—	Believing in authority	.44
IV	1	(7)	Maladjusted	—	Well-adjusted	.52
	2	(13)	Subjective	—	Objective	.52
	3	(14)	Anxious	—	Confident	.48
	4	(12)	Gloomy	—	Cheerful	.47
	5	(19)	Intolerant	—	Tolerant	.40

TABLE 57. *Clusters for examiner B*

Cluster			Variable			M_r
I	1	(14)	Anxious	—	Confident	.57
	2	(15)	Unsociable	—	Sociable	.55
	3	(6)	Dull	—	Acute	.53
	4	(18)	Lacking vitality	—	Vital	.50
	5	(11)	Dependent	—	Independent	.50
	6	(7)	Maladjusted	—	Well-adjusted	.48
	7	(2)	Reticent	—	Ostentatious	.48
	8	(13)	Subjective	—	Objective	.46
	9	(12)	Gloomy	—	Cheerful	.45
	10	(10)	Insensitive	—	Sensitive	.30
II	1	(7)	Maladjusted	—	Well-adjusted	.67
	2	(13)	Subjective	—	Objective	.61
	3	(12)	Gloomy	—	Cheerful	.58
	4	(9)	Scrupulous	—	Unscrupulous	.55
	5	(19)	Intolerant	—	Tolerant	.51
	6	(14)	Anxious	—	Confident	.50
III	1	(1)	Extrapunitive	—	Intropunitive	.46
	2	(4)	Oppositional	—	Believing in authority	.44
	3	(17)	Cold	—	Warmhearted	.40
	4	(3)	Dominant	—	Submissive	.38

TABLE 58. *Clusters for examiner C*

Cluster		Variable		M_r
I	1 (7)	Maladjusted	— Well-adjusted	.65
	2 (18)	Lacking vitality	— Vital	.61
	3 (13)	Subjective	— Objective	.60
	4 (15)	Unsociable	— Sociable	.59
	5 (12)	Gloomy	— Cheerful	.56
	6 (17)	Cold	— Warmhearted	.56
	7 (10)	Insensitive	— Sensitive	.54
	8 (14)	Anxious	— Confident	.50
	9 (19)	Intolerant	— Tolerant	.49
	10 (6)	Dull	— Acute	.48
	11 (4)	Oppositional	— Believing in authority	.45
II	1 (6)	Dull	— Acute	.59
	2 (18)	Lacking vitality	— Vital	.58
	3 (7)	Maladjusted	— Well-adjusted	.57
	4 (13)	Subjective	— Objective	.56
	5 (14)	Anxious	— Confident	.56
	6 (15)	Unsociable	— Sociable	.54
	7 (12)	Gloomy	— Cheerful	.51
	8 (3)	Submissive	— Dominant	.47
	9 (2)	Reticent	— Ostentatious	.45
	10 (11)	Dependent	— Independent	.42
III	1 (17)	Cold	— Warmhearted	.53
	2 (1)	Extrapunitive	— Intropunitive	.52
	3 (10)	Insensitive	— Sensitive	.51
	4 (9)	Unscrupulous	— Scrupulous	.30

TABLE 59. *Clusters for examiner D*

Cluster			Variable			M_r
I	1	(17)	Cold	—	Warmhearted	.46
	2	(18)	Lacking vitality	—	Vital	.45
	3	(15)	Unsociable	—	Sociable	.43
	4	(14)	Anxious	—	Confident	.42
	5	(2)	Reticent	—	Ostentatious	.41
	6	(12)	Gloomy	—	Cheerful	.41
II	1	(6)	Dull	—	Acute	.50
	2	(11)	Dependent	—	Independent	.48
	3	(18)	Lacking vitality	—	Vital	.43
	4	(10)	Insensitive	—	Sensitive	.39
	5	(3)	Submissive	—	Dominant	.29
III	1	(7)	Maladjusted	—	Well-adjusted	.51
	2	(14)	Anxious	—	Confident	.50
	3	(12)	Gloomy	—	Cheerful	.49
	4	(13)	Subjective	—	Objective	.29
IV	1	(1)	Extrapunitive	—	Intropunitive	.37
	2	(4)	Oppositional	—	Believing in authority	.32
	3	(19)	Intolerant	—	Tolerant	.28
V	1	(18)	Lacking vitality	—	Vital	.46
	2	(17)	Cold	—	Warmhearted	.45
	3	(9)	Unscrupulous	—	Scrupulous	.38
	4	(10)	Insensitive	—	Sensitive	.36

APPENDIX B

1. *An example of the rating forms*

6. *Logical reasoning*

Ability to draw conclusions and reason logically varies greatly. Some people think clearly and logically, combine rapidly, and are able to advance their views correctly and lucidly.

Some people, however, find it difficult to think coherently and concistently, and cannot easily arrive at valid conclusions. Most of us are average in this respect.

1. Dull
   Thinks unclearly
   Combines badly

4. Average

7. Acute
   Draws valid conclusions
   Combines well

	1	2	3	4	5	6	7
Magnus Davidsson							
Sven Karlsson							
Bengt Hansson							
etc.							

2. *The complete definitions of the variables*

1. *Aggression*

There are people who easily lose their tempers and lay the blame on others. They do not hesitate to reproach other people and find fault with their actions.

Others do not get angry, but think the fault is in themselves. They usually agree that adverse criticism of their actions is justified. Most people have no decided leanings in either direction.

1. Does not get angry with others
   Thinks the fault is within himself

4. Average

7. Easily angry with others
   Becomes agressive

2. *Shyness*

Some people like to be the centre of attraction. They are not afraid of appearing in public, and like to attract attention.

Others are retiring and afraid of coming into public notice. They are very shy.

Most of us are somewhere between these extremes.

1. Reticent
   Willing to take a suberdinate place

4. Average

7. Ostentatious
   Likes to appear in public

3. *Dominance—Submission*

Many people like to dominate—to control the activities of others—they often endeavour to impose their views on other people, to make them change their minds.

Others, however, are submissive and willing to agree with other's views and wishes; they allow others to make decisions for them.

Most of us are neither particularly dominant nor submissive.

1. Submissive	4. Average	7. Dominant Persuasive Authoritative

4. *Attitude towards superiors*

As we know, attitude towards authority is different in different people. Some are always in opposition to authority (e.g. parents, teachers or officers). They cannot subordinate themselves without protest.

Their opposites are eager to obey and resemble an authority, and admire, perhaps uncritically, a superior.

Neither of these descriptions suits most of us: we are somewhere in between these extremes.

1. Looks up to and admires uncritically a person in authority	4. Average	7. Oppositional Defiant Will not submit to authority

5. *Impulsive—Cautious*

There are people who are impulsive, and who say and do things on the spur of the moment, without thinking.

Others always think carefully before they act; they do nothing without thinking it over for a long time.

Most people are neither particularly impulsive nor cautious.

1. Cautious	4. Average	7. Impulsive Acts on the spur of the moment

6. *Logical reasoning*

Ability to draw conclusions and reason logically varies greatly. Some people think clearly and logically, combine rapidly, and are able to advance their views correctly and lucidly.

Some people, however, find it difficult to think coherently and consistently, and cannot easily arrive at valid conclusions.

Most of us are average in this respect.

1. Dull	4. Average	7. Acute
Thinks unclearly		Draws valid conclusions
Combines badly		Combines well

7. *Inner conflicts*

Some people seem to have great personal problems and conflicts. They often find it difficult to fit into their environment.

Others seem very harmonious and at home in most situations. They do not appear to have any great personal problems.

Most of us are somewhere between these two extremes. We have personal problems at times, but are well-balanced on the whole.

1. Well-adjusted	4. Average	7. Maladjusted
Well-balanced		Great personal problems

8. *Self-discipline*

Some people are very systematic and conscientious. They are very scrupulous, almost pedantic, even in details.

Others, however, are negligent, cannot take care of their belongings, are careless and forgetful.

Most people are neither particularly scrupulous nor negligent.

1. Negligent	4. Average	7. Careful
Careless		Pedantic

9. *Conscience*

Different people make different demands on themselves. Some make great demands on themselves, and cannot easily forgive themselves if they fail to fulfill these demands. There are many things they cannot forgive themselves.

Others allow themselves great freedom; they have no strict rules of behavior.

None of these extremes are applicable to most of us.

1. Allows themselves great freedom	4. Average	7. Makes great demands on themselves
Unscrupulous		Has strong principles
		Scrupulous

10. *Sensitivity*

There are people who are very sensitive to moods, changes and nuances. They may have a well-developed feeling for nature, and a delicate appreciation of art, music and poetry.

Others do not notice shades and changes.

These extremes are not applicable to most of us; we come about half way between.

1. Insensitive	4. Average	7. Sensitive
Does not notice shades and changes		Very sensitive to moods, changes and nuances

11. *Independent—Dependent*

There are people who have a both independent and matter-of-fact attitude, i.e. they are not afraid to be themselves and have an opinion of their own. They are not afraid to differ, without being obstinate and headstrong.

Their opposites are dependent upon others. They accommodate their views to those of their surroundings and are easily influenced by others.

In most cases we are neither particularly independent or dependent.

1. Very dependent	4. Average	7. Very independent
Easily influenced		Has an opinion of their own

12. *Disposition*

Some people are always cheerful and in a good temper; they are optimistic, and always look on the bright side.

Their opposites are often gloomy, depressed and melancholy.

Most people are neither very cheerful nor miserable.

1. Gloomy	4. Average	7. Very cheerful
Depressed		Optimistic

13. *Subjective—Objective*

Some people are inclined to take everything personally, they feel guilty, cannot distinguish between subject and person. They often misconstrue others' motives, and think that other people are talking about them.

Others judge their fellow-beings and their actions—like themselves—impartially and pertinently. It is easy for them to distinguish between subject and person.

Most of us have no definite leanings in either direction.

1. Objective	4. Average	7. Subjective
Impartial		Takes things personally

14. *Confident—Anxious*

There are people who always feel uncertain and anxious. They are afraid

of failure or of looking silly. They are very nervous when face to face with tests or new situations.

There are, on the other hand, those who are confident and unmoved in almost all situations. They are not afraid of criticism or of admitting failure.

Most of us are neither particularly confident nor anxious.

1. Confident Dares admit failure	4. Average	7. Very anxious and uncertain

15. *Sociable—Unsociable*

Many people like to be together with others, and come easily into contact with people. They are ingenuous and affable.

Others prefer to be alone; they are shy and reserved.

Most people are neither very sociable nor reserved.

1. Prefers to be alone Shy and reserved	4. Average	7. Very sociable, ingenuous and affable

16. *Reflection*

Some people like to sit and meditate, and they prefer to occupy themselves with theoretical rather than practical tasks. They often wonder what they are really like, and ponder over life.

Others are quite different, they do not like sitting and meditating. They do not worry so much about themselves and existence.

Most of us are about half way between these extremes.

1. Will not sit and meditate Does not like theoretical problems	4. Average	7. Likes very much to sit and meditate Analyses and reflects

17. *Warmhearted*

Some people are generous and warmhearted, and are sympathetic towards the people around them.

On the other hand there are people who seem chilly, reserved and indifferent to others.

Most of us are neither one nor the other to a very marked degree.

1. Cold Reserved Indifferent	4. Average	7. Very warmhearted

18. *Vitality*

There are people who are bursting with energy, they are stimulated by many and difficult tasks and, in spite of the intensive work they do, they seldom seem tired.

Others seem listless, lacking in vitality, easily tired and unable to stand strain.

Most of us are average in this respect.

1. Lacking vitality Easily tired	4. Average	7. Vital Energetic

19. *Tolerance*

Some people have very decided views as to how one should behave, and what opinions are correct. They are quick to censure views and actions that do not conform to their own.

Others view leniently actions and views that are not in keeping with their own. They are very broad-minded and tolerant.

1. Broad-minded Tolerant	4. Average	7. Narrow-minded Intolerant

20. *Rational—Emotional*

Many people react emotionally in nearly all situations. It is easy to see that they have a strongly developed emotional nature.

Others, however, react almost exclusively rationally. One seldom notices that they have any feelings of any kind.

Most people do not go to either extreme.

1. Rational	4. Average	7. Emotional

Index of abbreviations

Acta Ps.	Acta Psychologica. The Hague
Am. J. Psych.	The American Journal of Psychiatry. Baltimore, Md.
Ann. Rev. Ps.	Annual Review of Psychology. Stanford, Cal.
Brit. J. Stat. Ps.	British Journal of Statistical Psychology. London
Genet. Ps. Monogr.	Genetic Psychology Monograps. Provincetown, Mass.
J. Appl. Ps.	Journal of Applied Psychology. Lancaster, Pa.
J. Cons. Ps.	Journal of Consulting Psychology. Colorado Spring, Cal.
J. Educ. Ps.	Journal of Educational Psychology. Baltimore, Md.
J. Pers.	Journal of Personality. Durham, N.C.
J. Proj. Techn.	Journal of Projective Techinques. Glendale, Cal.
Ps. Bull.	Psychological Bulletin. Washington D.C.
Ps. Monogr.	Psychological Monographs. Washington D.C.

References

Anderson, H. H. and Anderson, G. L. (Eds.) An introduction to projective techniques. NY 1951. See MacFarlane, J. W. and Tuddenham, R. D.

Aron, B. A manual for analysis of the Thematic Apperception Test. Berkeley Calif. 1949.

Bellak, L. The TAT and CAT in clinical use. NY 1954.

Björkman, M. Validitet och konstans. Några teoretiska synpunkter illustrerade med en experimentell undersökning. Unpublished. Stockholm 1953.

Brehmer, I. En undersökning angående färgpyramidtestets tillförlitlighet. Unpublished. Stockholm 1959.

Bruner, J. S. and Tagiuri, R. The perception of people in Lindzey, G. (Ed.) *Handbook of social psychology.* Cambridge, Mass. 1954.

Buros, O. K. (Ed.) Third Mental Measurement Yearbook. New Jersey 1949. See Rotter, J.

Carlsson, G. Social class, intelligence and the verbal factor. *Acta Ps.*, 11 (1955), 269–78.

Cattell, *R. B.* Description and measurement of personality. NY 1946.

— Personality and Motivation. Structure and Measurement. NY 1957.

Crandall, V. J. Induced frustration and punishment—reward expectancy in Thematic Apperception stories. *J. Cons. Ps.*, 15 (1951), 400–3.

Cronbach, L. J. Assessment of individual differences. *Ann. Rev. Ps.*, 7 (1956), 173–96.

— and Gleser, G. C. Assessing similarity between profiles. *Ps. Bull.*, 50 (1953), 456–73.

— and Meehl, P. E. Construct validity in psychological tests. *Ps. Bull.*, 52 (1955), 281–302.

Davids, A. Alienation, social apperception and ego structure. *J. Cons. Ps.*, 19 (1955), 21–27.

— Personality dispositions, word frequency and word association. *J. Pers.*, 24 (1956), 328–38.

— and Rosenblatt, D. Use of the TAT in assessment of the personality syndrome of alienation. *J. Proj. Techn.*, 22 (1958), 145–52.

Derestad, R. En undersökning av tillförlitligheten i bedömningar baserade på Sentence-completion. Unpublished, Stockholm 1958.

Dooher, M. J. (ed.) and Marting, E. (co-ed.) Selection of management personnel, vol. I. NY 1957. See Taylor, E. K. and Nevis, E. C.

Ebel, R. L. Estimation of the reliability of ratings. *Psykometrika*, 16, (1951), 407–24.

Edwards, A. L. Experimental design in psykological research. NY 1950.

— Statistical Methods for the behavioral sciences. NY 1954.

Ekman, G. Bedömning och värdering. *Praktisk Psykologi*, 2 (1946).
— Reliabilitet och konstans. Uppsala 1947.
— Subjektiv bedömning in Westerlund, G. Människan och arbetet. Stockholm 1951.
Eron, L. D. Some problems in the research application of the Thematic Apperception Test. *J. Proj. Techn.*, 19 (1955), 125–9.
— and Ritter, A. A comparison of two methods of administration of the Thematic Apperception Test. *J. Cons. Ps.*, 15 (1951), 55–61.
Eysenck, H. J. Criterion analysis—An application of the hypothetico-deductive method to factor analysis. *Psychological Review*, 1950, 38–53.
Fisher, R. A. and Yates, F. Statistical Tables for Biological, Agricultural and Medical Research. London 1953.
Folkhögskolan och de stora årskullarna. Förslag av Skolöverstyrelsens planeringskommitté för de stora årskullarna. Stencil. Unpublished, Stockholm 1956.
Fruchter, B. Introduction to factor analysis. NY 1954.
Garfield, S. L., Blek, L. and Melker, F. The influence of method of administration and sex differences on selected aspects of TAT stories. *J. Cons. Ps.*, 16 (1952), 140–44.
Guilford, J. P. Fundamental Statistics in Psychology and Education. NY 1950.
— Psychometric Methods. NY 1954.
— Personality. NY 1959.
Harrison, R. Studies in the use and validity of the Thematic Apperception Test with mentally disorded patients. *Character and Personality*, 9 (1940).
Hartman, A. A. An experimental examination of the Thematic Apperception technique in clinical diagnosis. *Ps. Monogr.*, 63 (1949).
Heiss, R. und Hiltmann, H. Der Farbpyramidentest. Bern 1951.
Henry, W. E. The Thematic Apperception technique in the study of culture-personality relations. *Genet. Ps. Monogr.*, 35 (1947), 3–135.
Henrysson, S. Applicability of factor analysis in the behavioral sciences. A methodological study. Uppsala 1957.
Holzinger, K. J. and Harman, H. H. Factor Analysis. A synthesis of factorial methods. Chicago 1941.
Härnqvist, K. et. al. Redogörelse för konstruktionsarbete med ett gruppintelligenstest. Unpublished, Stockholm 1956.
— Nyheter inom teorin för psykologisk och pedagogisk testning. En översikt. Göteborg 1959.
Johnsson, D. M. and Vidulich, R. N. Experimental manipulation of the halo effect. *J. Appl. Ps.*, 40 (1956), 130–5.
Jones, L. V. and Fiske, D. W. Models for testing the significance of combined results. *Ps. Bull.*, 50 (1953), 375–82.
Jonsson, C-O. Questionnaries and Interviews. Experimental studies concerning concurrent validity on well-motivated subjects. Stockholm 1957.
Kelly, E. L. The selection of upper and lower groups for the validation of test items. *J. Educ. Ps.*, 30 (1939), 17–24.

KELLY, E. L. and FISKE, D. W. The prediction of success in the VA-training program in clinical psychology. *The American Psychologist,* 5 (1950), 395–406.

— The prediction of performance in clinical psychology. *Ann Arbor,* Mich. 1951.

KLOPFER, B. and others. Developments in the Rorschach techniques. Vol. II. NY 1956. See SHNEIDMAN, E. S.

LINDBERG, B. J. Om innebörden av färgsvaren vid ringprovet. *Tidskr. för psyk. och ped.,* 1 (1943).

— Nya experiment till belysning av färginställningen vid ringprovet. *Tidskr. för psyk. och ped.,* 3–4. Särtryck 1949.

LINDQUIST, E. F. Statistical Analysis in Educational Research. Cambridge, Mass. 1940.

— Design and analysis of experiments in psychology and education. Boston 1956.

LINDZEY, G. (Ed.) Handbook of social psychology. Cambridge, Mass. 1954. See BRUNER, J. S. and TAGIURI, R.

— and HEINEMAN, S. H. Thematic Apperception Test: Individual and group administration. *J. Pers.,* 24 (1955), 34–55.

— and HERMAN, P. TAT: A note on reliability and situational validity. *J. Proj. Techn.,* 19 (1955), 36–42.

MACFARLANE, J. W. and TUDDENHAM, R. D. Problems in the validation of projective techniques in ANDERSON, H. H. and ANDERSON, G. L. (Eds.) An introduction to projective techniques. NY 1951.

MAGNUSSON, D. En undersökning av tillförlitligheten av vissa projektiva test, särskilt TAT. Unpublished, Stockholm 1957.

MCNEMAR, Q. Psychological statistics. Second ed. NY 1955.

MILAM, J. Examiner influences on TAT-stories. *J. Proj. Techn.,* 18 (1954), 221–6.

MORE, D. M. The congruence of projective instruments in personnel assessment. *J. Appl. Psych.,* 41 (1957), 137–40.

MURRAY, H. A. Explorations in personality. NY 1938.

— et.al. Thematic Apperception Test Manual. Cambridge 1943.

— Uses of the Thematic Apperception Test. *Am. J. Psych.,* 107 (1950/1951).

PETERS, C. C. and VAN VOORHIS, W. R. Statistical procedures and their mathematical bases. NY 1940.

ROTTER, J. The Thematic Apperception Test in BUROS, O. K. (Ed.) *Third Mental Measurement Yearbook.* New Jersey 1949.

— Clinical methods: Psychodiagnostics. *Ann. Rev. Ps.* 1953, 295–316.

SANFORD, R. N., ADKINS, M. M., MILLER, R. B. and COBB, E. A. Physique, Personality and Scholarship. A cooperative study of school children. Washington 1943.

SEN, A. A preliminary study of the TAT. *Brit. J. Stat. Ps.,* 6 (1953), 91–100.

SHATIN, L. Rorschach adjustment and the Thematic Apperception Test. *J. Proj. Techn.,* 17 (1953), 92–101.

— Relationships between the Rorschach test and the Thematic Apperception Test. *J. Proj. Techn.*, 19 (1955), 317–31.

SHNEIDMAN, E. S. Some relationships between the Rorschach technique and other psychodiagnostic tests in KLOPFER, B. and others, Developments in the Rorschach techniques. Vol. II. NY 1956.

SIEGEL, S. Nonparametric statistics. For the behavioral sciences. NY 1956.

STEIN, M. I. The Thematic Apperception Test. An introductory manual for the clinical use for adults. Cambridge, Mass. 1955.

TAYLOR, E. K. and NEVIS, E. C. The use of projective techniques at management levels in DOOHER, M. J. (ed.) and MARTING, E. (co-ed.). Selection of management personnel. NY 1957. Vol. I.

Technical recommendations for psychological tests and diagnostic techniques. *Ps. Bull. Suppl.*, 51 (1954), 1–38.

The Royal Board of Education. Survey of the school system in Sweden. Stockholm 1958.

THURSTONE, L. L. Multiple-factor analysis. Chicago 1947.

TOMKINS, S. S. The Thematic Apperception Test. The theory and technique of interpretation. NY 1947.

VERNON, PH. E. Personality tests and assessments. London 1957.

WEIDER, A., WOLFF, H., BRODMAN, K., MITTELMAN, B. and WECHSLER, D. Cornell Index. NY 1949.

WESTERLUND, G. (Ed.) Människan och arbetet. Stockholm 1951. See EKMAN, G.

WITTENBORN, J. R. The theory and technique of assessment. *Ann. Rev. Ps.* 1957, 331–56.

ÅS, A. Mutilation Fantasies and Autonomic Response. Oslo 1958.

Published by
THE SWEDISH COUNCIL FOR
PERSONELL ADMINISTRATION
HANTVERKARGATAN 78
STOCKHOLM K — SWEDEN

Price 26: — Swedish Crowns
International distribution by
ALMQVIST & WIKSELL, Stockholm C, Sweden